Published in Journals: AI, Big Data and Cognitive Computing, International Journal of Financial Studies, Journal of Theoretical and Applied Electronic Commerce Research and Sustainability

Topic Reprint

Artificial Intelligence Applications in Financial Technology

Edited by
Albert Y.S. Lam and Yanhui Geng

mdpi.com/topics

MDPI

Artificial Intelligence Applications in Financial Technology

Artificial Intelligence Applications in Financial Technology

Topic Editors

Albert Y.S. Lam
Yanhui Geng

MDPI

Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Topic Editors

Albert Y.S. Lam
Fano Labs and Department of Electrical and Electronic Engineering
The University of Hong Kong
Hong Kong

Yanhui Geng
Huawei Hong Kong Research Centre
Hutchison Telecom Tower
Hong Kong

Editorial Office
MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of the Topic, published open access by the journals *AI* (ISSN 2673-2688), *Big Data and Cognitive Computing* (ISSN 2504-2289), *International Journal of Financial Studies* (ISSN 2227-7072), *Journal of Theoretical and Applied Electronic Commerce Research* (ISSN 0718-1876) and *Sustainability* (ISSN 2071-1050), freely accessible at: https://www.mdpi.com/topics/AI_financial.

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

Lastname, A.A.; Lastname, B.B. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-7258-5183-6 (Hbk)
ISBN 978-3-7258-5184-3 (PDF)
https://doi.org/10.3390/books978-3-7258-5184-3

© 2025 by the authors. Articles in this book are Open Access and distributed under the Creative Commons Attribution (CC BY) license. The book as a whole is distributed by MDPI under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) license (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

About the Editors . . . . . **ix**

Albert Y. S. Lam
Artificial Intelligence Applications in Financial Technology
Reprinted from: *J. Theor. Appl. Electron. Commer. Res.* **2025**, *20*, 29,
https://doi.org/10.3390/jtaer20010029 . . . . . **1**

Dana AL-Najjar, Nadia Al-Rousan and Hazem AL-Najjar
Machine Learning to Develop Credit Card Customer Churn Prediction
Reprinted from: *J. Theor. Appl. Electron. Commer. Res.* **2022**, *17*, 1529-1542,
https://doi.org/10.3390/jtaer17040077 . . . . . **6**

Ying Liu, Haoran Zhao, Jieguang Sun and Yahui Tang
Digital Inclusive Finance and Family Wealth: Evidence from LightGBM Approach
Reprinted from: *Sustainability* **2022**, *14*, 15363,
https://doi.org/10.3390/su142215363 . . . . . **20**

Yasheng Chen, Zhuojun Wu and Hui Yan
A Full Population Auditing Method Based on Machine Learning
Reprinted from: *Sustainability* **2022**, *14*, 17008,
https://doi.org/10.3390/su142417008 . . . . . **39**

Yasheng Chen and Zhuojun Wu
Financial Fraud Detection of Listed Companies in China: A Machine Learning Approach
Reprinted from: *Sustainability* **2023**, *15*, 105,
https://doi.org/10.3390/su15010105 . . . . . **56**

Adrian Millea and Abbas Edalat
Using Deep Reinforcement Learning with Hierarchical Risk Parity for Portfolio Optimization
Reprinted from: *Int. J. Financial Stud.* **2023**, *11*, 10,
https://doi.org/10.3390/ijfs11010010 . . . . . **71**

Amal Al Ali, Ahmed M. Khedr, Magdi El Bannany and Sakeena Kanakkayil
GALSTM-FDP: A Time-Series Modeling Approach Using Hybrid GA and LSTM for Financial Distress Prediction
Reprinted from: *Int. J. Financial Stud.* **2023**, *11*, 38,
https://doi.org/10.3390/ijfs11010038 . . . . . **87**

Seol-Hyun Noh
Comparing the Performance of Corporate Bankruptcy Prediction Models Based on Imbalanced Financial Data
Reprinted from: *Sustainability* **2023**, *15*, 4794,
https://doi.org/10.3390/su15064794 . . . . . **102**

Wentao Hu and Xiaoxiao Li
Financial Technology Development and Green Total Factor Productivity
Reprinted from: *Sustainability* **2023**, *15*, 10309,
https://doi.org/10.3390/su151310309 . . . . . **119**

Jorge de Andrés-Sánchez and Jaume Gené-Albesa
Explaining Policyholders' Chatbot Acceptance with an Unified Technology Acceptance and Use of Technology-Based Model
Reprinted from: *J. Theor. Appl. Electron. Commer. Res.* **2023**, *18*, 1217-1237,
https://doi.org/10.3390/jtaer18030062 . . . . . **147**

Markus Frohmann, Manuel Karner, Said Khudoyan, Robert Wagner and Markus Schedl
Predicting the Price of Bitcoin Using Sentiment-Enriched Time Series Forecasting
Reprinted from: *Big Data Cogn. Comput.* **2023**, *7*, 137,
https://doi.org/10.3390/bdcc7030137 . . . . . **168**

Chew Yee Patience Cheah, Yue Yang and Boon Giin Lee
Enhancing Financial Fraud Detection through Addressing Class Imbalance Using Hybrid SMOTE-GAN Techniques
Reprinted from: *Int. J. Financial Stud.* **2023**, *11*, 110,
https://doi.org/10.3390/ijfs11030110 . . . . . **188**

Fadi Shehab Shiyyab, Abdallah Bader Alzoubi, Qais Mohammad Obidat and Hashem Alshurafat
The Impact of Artificial Intelligence Disclosure on Financial Performance
Reprinted from: *Int. J. Financial Stud.* **2023**, *11*, 115,
https://doi.org/10.3390/ijfs11030115 . . . . . **205**

I. de Zarzà, J. de Curtò, Gemma Roig and Carlos T. Calafate
Optimized Financial Planning: Integrating Individual and Cooperative Budgeting Models with LLM Recommendations
Reprinted from: *AI* **2024**, *5*, 6,
https://doi.org/10.3390/ai5010006 . . . . . **230**

Bassant A. Abdelfattah, Saad M. Darwish and Saleh M. Elkaffas
Enhancing the Prediction of Stock Market Movement Using Neutrosophic-Logic-Based Sentiment Analysis
Reprinted from: *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 116-134,
https://doi.org/10.3390/jtaer19010007 . . . . . **254**

Sizheng Wei and Suan Lee
Financial Anti-Fraud Based on Dual-Channel Graph Attention Network
Reprinted from: *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 16,
https://doi.org/10.3390/jtaer19010016 . . . . . **273**

Chiou-Yann Lee, Chun-Ru Wen and Binh Thi-Thanh-Nguyen
Board Expertise Background and Firm Performance
Reprinted from: *Int. J. Financial Stud.* **2024**, *12*, 17,
https://doi.org/10.3390/ijfs12010017 . . . . . **291**

Anuwat Boonprasope and Korrakot Yaibuathet Tippayawong
Predicting Healthcare Mutual Fund Performance Using Deep Learning and Linear Regression
Reprinted from: *Int. J. Financial Stud.* **2024**, *12*, 23,
https://doi.org/10.3390/ijfs12010023 . . . . . **306**

Adela Bâra and Simona-Vasilica Oprea
The Impact of Academic Publications over the Last Decade on Historical Bitcoin Prices Using Generative Models
Reprinted from: *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 538-560,
https://doi.org/10.3390/jtaer19010029 . . . . . **327**

Hadi Gholampoor and Majid Asadi
Risk Analysis of Bankruptcy in the U.S. Healthcare Industries Based on Financial Ratios: A Machine Learning Analysis
Reprinted from: *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 1303-1320,
https://doi.org/10.3390/jtaer19020066 . . . . . **350**

Tianyi Cao, Xinrui Wan, Huanhuan Wang, Xin Yu and Libo Xu
Quantitative Stock Selection Model Using Graph Learning and a Spatial–Temporal Encoder
Reprinted from: *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 1756-1775,
https://doi.org/10.3390/jtaer19030086 . . . . . **368**

About the Editors

Albert Y.S. Lam

Albert Y.S. Lam received a BEng degree (First Class Honors) in Information Engineering from the University of Hong Kong (HKU), Hong Kong, in 2005, and he obtained a PhD degree at the Department of Electrical and Electronic Engineering of HKU in 2010. He was a postdoctoral scholar at the Department of Electrical Engineering and Computer Sciences of University of California, Berkeley. He was a Research Assistant Professor at the Department of Computer Science of Hong Kong Baptist University in 2012–15 and the Department of Electrical and Electronic Engineering (EEE) of HKU in 2015–17. He is now the Chief Technology Officer and Chief Scientist at Fano Labs, a deep-tech startup specializing in speech and language technologies. He also serves as an Adjunct Assistant Professor at HKU. He is a Croucher research fellow. He is one of the top 2% scientists Worldwide according to Stanford University, 2020–24, and is an active member of IEEE. He is the founding Chair of the Social Media Subcommittee of the IEEE Computational Intelligence Society (CIS) and has also chaired some other committees in CIS. His research interests include optimization theory and algorithms, artificial intelligence, evolutionary computation, smart grids, and smart cities. He is Senior Editor of *IEEE Transactions on Intelligent Transportation Systems* and an Associate Editor of *IEEE Transactions on Evolutionary Computation*, *IEEE Transactions on Artificial Intelligence*, and *IEEE Transactions on Emerging Topics in Computational Intelligence*. He is also an Editor-in-Chief of *EAI Endorsed Transactions on Energy Web*.

Yanhui Geng

Yanhui Geng received his BEng and MEng degrees in Electronic Engineering and Information Science from the University of Science and Technology of China (USTC), Hefei, China, in 2002 and 2005, respectively, and a Ph.D. degree in Electrical and Electronic Engineering from The University of Hong Kong (HKU), Hong Kong, in 2009. He is currently the Head of the Huawei Hong Kong Research Center, Hong Kong. Before that, he was the Head of the Huawei Montreal Research Center from 2017 to 2020 and was a Principal Researcher at Huawei Noah's Ark Lab, Hong Kong, from 2013 to 2017. Before joining Huawei, he was a postdoctoral research fellow at HKU from 2009 to 2012 and a senior scientific engineer at the Hong Kong Applied Science and Technology Research Institute (ASTRI), Hong Kong, from 2012 to 2013. He has filed over 20 patents globally and he has over 30 publications in top journals and conferences. His research interests include artificial intelligence, machine learning, information theory, wireless communication, and datacenter networking. He received numerous awards, including the IEEE ICC 2010 Best Paper Award, the Chinese Academy of Science President Award, the Excellent Postgraduate of Anhui Province Award, and the Excellent Postgraduate of USTC Award.

Journal of
Theoretical and Applied
Electronic Commerce Research

MDPI

Editorial

Artificial Intelligence Applications in Financial Technology

Albert Y. S. Lam

Fano AI, Hong Kong, China; albert@fano.ai

1. Introduction

Financial technology, commonly known as fintech, represents the intersection of finance and information technology aimed at simplifying, enhancing, transforming, and automating financial processes and services for businesses and individuals. Traditionally, the financial industry has relied heavily on human intervention, which often leads to high costs, inefficiencies, errors, compliance issues, and penalty fines. These processes typically involve extensive document handling and complex communications between various stakeholders, such as agents and customers, supervisors and subordinates, and institutions and regulators. However, fintech leverages electronic devices and advanced technologies to streamline these processes, ensuring 24/7 operation with consistent efficiency and accuracy. The integration of artificial intelligence (AI) within fintech has been particularly transformative, as AI endows machines with cognitive abilities akin to human intelligence, enabling the automation of tasks that were once thought to be the exclusive domain of humans. This includes tasks related to image recognition, natural language processing (NLP), and speech recognition. For instance, AI can convert handwritten documents or printed text into electronic formats, making it easier to analyze and store information. NLP facilitates the extraction of valuable information from text, while speech recognition technology enables the analysis of verbal conversations, thus enhancing the overall efficiency and effectiveness of financial operations.

The application of AI in fintech has revolutionized the global banking, financial services, and insurance (BFSI) industry. AI-driven tools and solutions have been developed to address specific needs within this industry, leading to the emergence of specialized disciplines such as regtech, suptech, and insurtech. Regtech focuses on the management of regulatory processes, ensuring that financial institutions comply with ever-evolving regulations. Suptech, on the other hand, is concerned with regulatory supervision and oversight, utilizing AI to monitor and ensure compliance within the financial sector. Insurtech leverages AI to design innovative insurance products and solutions, enhancing the overall customer experience and operational efficiency. Furthermore, AI has the potential to revolutionize various aspects of financial processes, making them more efficient and error-free. One area where AI has made a significant impact is in fraud detection and prevention. Traditional methods of fraud detection often involve manual review processes, which are time-consuming and prone to errors. AI, with its ability to analyze vast amounts of data quickly and accurately, can identify suspicious activities and patterns that may indicate fraudulent behavior. Machine learning algorithms can be trained to recognize anomalies and flag potential fraud in real-time, reducing the risk of financial losses and safeguarding the integrity of financial transactions.

AI is also playing a crucial role in automating various financial services, from loan processing to investment management. In the lending industry, AI-powered algorithms can assess the creditworthiness of applicants more accurately and efficiently than traditional

methods. By analyzing a wide range of data points, ranging from financial history and employment records to social media activity, AI can provide a comprehensive risk assessment and make faster lending decisions. This not only speeds up the loan approval process, but also reduces the likelihood of defaults, thereby benefiting both lenders and borrowers. In the realm of investment management, AI-driven robo-advisors are gaining popularity. These automated platforms use sophisticated algorithms to analyze market trends, assess risk tolerance, and recommend investment strategies tailored to individual investors. Robo-advisors can continuously monitor and adjust investment portfolios based on real-time data, ensuring optimal performance and minimizing risks. This democratizes access to investment advice, making it available to a broader audience at a lower cost compared to traditional financial advisors. Additionally, regulatory compliance is a critical aspect of the financial industry, and AI is proving to be an invaluable tool in this regard. By automating compliance checks and monitoring regulatory changes, AI helps financial institutions stay up-to-date with the latest requirements, reducing the risk of non-compliance and associated penalties. Overall, the integration of AI in fintech is driving significant improvements in efficiency, accuracy, and customer satisfaction, making it an indispensable component of the modern financial landscape.

Although AI is promising in fintech, the practitioners are struggling with its applications due to the following challenges:

- Data quality and bias: High-quality data are necessary for effective AI in fintech. Bad data may produce poor predictions and undependable insights, while biases lead to discriminatory practices in the likes of biased credit scoring. The biases usually come from historical data reflecting inequalities in society. Ensuring nondiscriminatory data can be achieved through cleaning and validation, and advanced techniques to find and reduce bias.
- Ethical AI practices: The AI ethics in fintech is related to its fairness, transparency, and accountability. Fairness can be considered as the avoidance of discrimination, transparency can be promoted by explainability, and responsibility by accountability. The development of ethical practices requires continuous training, clear guidelines, and control mechanisms. Certain ethical dilemmas, such as those in automated trading affecting market stability, should be carefully handled.
- Cybersecurity vulnerabilities: Integration of AI in fintech involves cybersecurity issues, as financial AI systems can be targeted by cyberattacks relatively easily or create vulnerabilities. This is a concern for adversarial attacks, where malicious inputs deceive the system. To protect AI systems, robust cybersecurity measures are required, including encryption, continuous monitoring, and secure development practices. As AI complexity increases, sophisticated security measures become vital to safeguard financial data and ensure system integrity.
- Integration with legacy systems: Many financial institutions still operate with legacy systems that are inefficient for the integration of modern AI technologies due to technological complications and its high cost. These legacy systems often lack necessary infrastructure and flexibility, and thus substantial investment is required to upgrade or overhaul them. Overcoming such integration challenges will effectively guarantee maximum utilization of AI in fintech for increased efficiency and innovative financial services.

We are at a time when AI technologies evolve incredibly and impact our daily lives incrementally. AI will revolutionize the financial sector with more personalization, better security, and enhanced automation. For example, virtual assistants and chatbots will continue to give personalized financial advice while predictive analytics will increasingly inform customer expectations. Real-time anomaly detection and biometric authentication

will move the needle in the right direction through fraud detection and cybersecurity. Further down the road, advances in algorithmic trading and robo-advisors rest on the principles of machine learning being deployed further in market analysis and investment management. Finally, AI will transform credit scoring while using alternative data sources, enabling financial inclusion for marginalized populations.

This Special Topic contributes to the fintech agenda through enhanced scientific and multi-disciplinary knowledge intended to improve the performance and deployment by bringing focus to AI, in order to meet technical, economical, and social goals. We are particularly interested in understand how various AI technologies can be adopted in the financial world intuitively and practically.

2. Summary of the Contributions

Research works by scholars from divergent backgrounds are included in this Special Issue. They can be classified into four themes, namely: (i) machine learning and AI applications in financial prediction and risk management, (ii) the impact of AI on financial and organizational performance, (iii) predicting financial asset prices using sentiment analysis and machine learning, and (iv) advanced machine learning techniques for financial fraud detection.

(i) Refs. [1–8] explore various applications of machine learning and deep learning techniques in financial contexts, including predicting credit card customer churn, household wealth, financial distress, corporate bankruptcy, healthcare industry bankruptcy, stock selection, and portfolio optimization, demonstrating improvements in predictive accuracy and decision-making processes. In [1], a feature-election method and several machine learning models are adopted to develop a predictive model, which forecasts customer status changes for potential churns. Ref. [2] designs a predictive model based on the LightGBM machine learning algorithm to predict household wealth and to analyze the features of strong predictive ability for household wealth. While the traditional audit methods are based on sampling, ref. [3] proposes a machine learning-based full population auditing method to cover all audit samples through self-learning feature. In [4], a hybrid approach based on genetic algorithm and long short-term memory (LSTM) is adopted to build a time-series model to predict financial distress. Ref. [5] proposes a method to process corporate imbalanced financial data for corporate bankruptcy prediction. Multi-class classification models are developed in [6] to predict bankruptcy risk in the U.S. healthcare sector. Ref. [7] makes use of a graph learning spatial-temporal encoder network for quantitative stock selection to withstand market turbulence and enhance profitability. Ref. [8] designs a hierarchical decision-marking architecture for portfolio optimization to leverage the flexibility of deep reinforcement learning and the reliability and efficiency of the hierarchical risk models.

(ii) We explore the impact of fintech on various aspects of financial performance, such as productivity, AI disclosure, and the acceptance of new technologies in financial services like chatbots. In [9], a slack-based measure model is utilized to measure the green total factor productivity (GTFP) levels of cities in China and to empirically test the impact of fintech on GTFP and the underlying mechanisms. Ref. [10] studies how Jordanian banks make use of AI technologies in their operations and evaluates the impact of AI disclosure on financial performance. While AI-based conversational robots are increasing to be adopted for customer facing in the insurance industry, ref. [11] investigates the level of consumers' acceptance to interaction with AI and identifies various influential factors for behavioral intention. By integrating with large language model recommendation, ref. [12] designs an optimization framework

for individual budget allocation and for household for handling multiple incomes and shared expenses. In [13], a financial performance forecasting method, based on threshold technique and artificial neural network, is developed to identify factors from the board of directors impacting the financial performance of various industries in Taiwan.

(iii) Refs. [14–17] investigate advanced techniques in predicting financial asset prices, such as Bitcoin and stock market movements, using sentiment analysis, deep learning, and generative models, emphasizing the integration of external factors and novel algorithms to improve prediction accuracy and robustness. In [14], a hybrid method, which combines sentiment prediction from microblogs and time series forecasting, is developed to predict the intraday price of Bitcoin. Ref. [15] considers a different approach, in which BERT and Word2Vec are used to analyze the impact of academic publications on Bitcoin prices. In [16], neutrosophic logic and LSTM are utilized for sentiment analysis to investigate the impact of social media on stock market. Ref. [17] forecasts the post-COVID price trends of mutual funds in healthcare by using multiple linear regression and LSTM.

(iv) Refs. [18–20] explore advanced machine learning techniques to detect financial fraud, emphasizing the effectiveness of ensemble learning algorithms, hybrid data generation methods, and graph neural networks combined with blockchain technology to improve fraud detection accuracy and data security. Ref. [18] focuses on fraud identification through the structured data in financial reports to improve audit efficiency and reduce the dependence on non-financial data. Ref. [19] proposes to use the synthetic minority oversampling technique and generative adversarial network to address the class imbalance problem found in finance fraud datasets. In [20], blockchain technology is incorporated with graph-based neural network to enhance fraud identification in financial transactions while preserving user privacy.

Funding: This research received no external funding.

Conflicts of Interest: Author Albert Y. S. Lam was employed by the company Fano AI. The author declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Al-Najjar, D.; Al-Najjar, N.; Al-Najjar, H. Machine Learning to Develop Credit Card Customer Churn Prediction. *J. Theor. Appl. Electron. Commer. Res.* **2022**, *17*, 1529–1542. [CrossRef]
2. Liu, Y.; Zhao, H.; Sun, J.; Tang, Y. Digital Inclusive Finance and Family Wealth: Evidence from LightGBM Approach. *Sustainability* **2022**, *14*, 15363. [CrossRef]
3. Chen, Y.; Wu, Z.; Yan, H. A Full Population Auditing Method Based on Machine Learning. *Sustainability* **2022**, *14*, 17008. [CrossRef]
4. Al Ali, A.; Khedr, A.M.; El Bannany, M.; Kanakkayil, S. GALSTM-FDP: A Time-Series Modeling Approach Using Hybrid GA and LSTM for Financial Distress Prediction. *Int. J. Financ. Stud.* **2023**, *11*, 38. [CrossRef]
5. Noh, S.-H. Comparing the Performance of Corporate Bankruptcy Prediction Models Based on Imbalanced Financial Data. *Sustainability* **2023**, *15*, 4794. [CrossRef]
6. Gholampoor, H.; Asadi, M. Risk Analysis of Bankruptcy in the U.S. Healthcare Industries Based on Financial Ratios: A Machine Learning Analysis. *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 1303–1320. [CrossRef]
7. Cao, T.; Wan, X.; Wang, H.; Yu, X.; Xu, L. Quantitative Stock Selection Model Using Graph Learning and a Spatial–Temporal Encoder. *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 1756–1775. [CrossRef]
8. Millea, A.; Edalat, A. Using Deep Reinforcement Learning with Hierarchical Risk Parity for Portfolio Optimization. *Int. J. Financ. Stud.* **2023**, *11*, 10. [CrossRef]
9. Hu, W.; Li, X. Financial Technology Development and Green Total Factor Productivity. *Sustainability* **2023**, *15*, 10309. [CrossRef]
10. Shiyyab, F.S.; Alzoubi, A.B.; Obidat, Q.M.; Alshurafat, H. The Impact of Artificial Intelligence Disclosure on Financial Performance. *Int. J. Financ. Stud.* **2023**, *11*, 115. [CrossRef]

11. de Andrés-Sánchez, J.; Gené-Albesa, J. Explaining Policyholders' Chatbot Acceptance with an Unified Technology Acceptance and Use of Technology-Based Model. *J. Theor. Appl. Electron. Commer. Res.* **2023**, *18*, 1217–1237. [CrossRef]
12. de Zarzà, I.; de Curtò, J.; Roig, G.; Calafate, C.T. Optimized Financial Planning: Integrating Individual and Cooperative Budgeting Models with LLM Recommendations. *AI* **2024**, *5*, 91–114. [CrossRef]
13. Lee, C.-Y.; Wen, C.-R.; Thi-Thanh-Nguyen, B. Board Expertise Background and Firm Performance. *Int. J. Financ. Stud.* **2024**, *12*, 17. [CrossRef]
14. Frohmann, M.; Karner, M.; Khudoyan, S.; Wagner, R.; Schedl, M. Predicting the Price of Bitcoin Using Sentiment-Enriched Time Series Forecasting. *Big Data Cogn. Comput.* **2023**, *7*, 137. [CrossRef]
15. Bâra, A.; Oprea, S.-V. The Impact of Academic Publications over the Last Decade on Historical Bitcoin Prices Using Generative Models. *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 538–560. [CrossRef]
16. Abdelfattah, B.A.; Darwish, S.M.; Elkaffas, S.M. Enhancing the Prediction of Stock Market Movement Using Neutrosophic-Logic-Based Sentiment Analysis. *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 116–134. [CrossRef]
17. Boonprasope, A.; Tippayawong, K.Y. Predicting Healthcare Mutual Fund Performance Using Deep Learning and Linear Regression. *Int. J. Financial Stud.* **2024**, *12*, 23. [CrossRef]
18. Chen, Y.; Wu, Z. Financial Fraud Detection of Listed Companies in China: A Machine Learning Approach. *Sustainability* **2023**, *15*, 105. [CrossRef]
19. Cheah, P.C.Y.; Yang, Y.; Lee, B.G. Enhancing Financial Fraud Detection through Addressing Class Imbalance Using Hybrid SMOTE-GAN Techniques. *Int. J. Financ. Stud.* **2023**, *11*, 110. [CrossRef]
20. Wei, S.; Lee, S. Financial Anti-Fraud Based on Dual-Channel Graph Attention Network. *J. Theor. Appl. Electron. Commer.* **2024**, *19*, 297–314. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Journal of
Theoretical and Applied
Electronic Commerce Research

MDPI

Article

Machine Learning to Develop Credit Card Customer Churn Prediction

Dana AL-Najjar [1], Nadia Al-Rousan [2] and Hazem AL-Najjar [3,*]

[1] Department of Finance and Banking Sciences, Faculty of Business, Applied Science Private University, Amman 11931, Jordan
[2] MIS Department, Faculty of Business, Sohar University, Sohar 311, Oman
[3] Department of Computer Engineering, Faculty of Engineering and Architecture, Istanbul Gelisim University, 34310 Istanbul, Turkey
* Correspondence: hazem_najjar@yahoo.com

Abstract: The credit card customer churn rate is the percentage of a bank's customers that stop using that bank's services. Hence, developing a prediction model to predict the expected status for the customers will generate an early alert for banks to change the service for that customer or to offer them new services. This paper aims to develop credit card customer churn prediction by using a feature-selection method and five machine learning models. To select the independent variables, three models were used, including selection of all independent variables, two-step clustering and k-nearest neighbor, and feature selection. In addition, five machine learning prediction models were selected, including the Bayesian network, the C5 tree, the chi-square automatic interaction detection (CHAID) tree, the classification and regression (CR) tree, and a neural network. The analysis showed that all the machine learning models could predict the credit card customer churn model. In addition, the results showed that the C5 tree machine learning model performed the best in comparison with the three developed models. The results indicated that the top three variables needed in the development of the C5 tree customer churn prediction model were the total transaction count, the total revolving balance on the credit card, and the change in the transaction count. Finally, the results revealed that merging the multi-categorical variables into one variable improved the performance of the prediction models.

Keywords: customer churn; machine learning; feature selection; two-step clustering; prediction model

1. Introduction

At present, the market is dynamic and highly competitive due to the availability of large numbers of service providers, especially banks, worldwide. One of the main challenges for this sector is the change in customer behavior. Customers are the core of all industries, especially customer-dependent organizations, such as the banking sector, which is responsible for accepting deposits, making investments, and granting loans. Long-term customers are directly connected to the production of profits; hence, banks should avoid losing customers [1–4]. The *Harvard Business Review* believes that a 5% defection in customers can lead to an increase in profits for firms of between 25% and 85% [4,5]. Thus, given that customers are the most important assets with strong effects on a bank's profit, there are five essential pillars for the modern banking business: capital, liquidity, risk, assets, and customer management [6,7]. Focusing effectively on the five pillars can ensure that management effectively maximizes the profits of a bank [7,8].

Therefore, customer churn is a fundamental challenge for banks. Customer churn can be defined as the loss of a customer to a competitor, which leads to losses in profits. To manage churning, it is essential to identify those customers who are likely to move to a competing bank [9–11]. In addition, Risselada et al. (2010) [12] showed that churning management is important in establishing appropriate long-term relationships between

firms and customers to maximize the value of the customer base. Customer churn can be divided into two groups: voluntary churn and nonvoluntary churn. Nonvoluntary churn occurs when the bank withdraws services from customers, and it is easy to detect. On the other hand, voluntary churn is more difficult to identify, because it is a conscious decision by a customer to terminate their relationship with a certain bank. Moreover, it can be subdivided into incidental churn and deliberate churn: incidental churn occurs when there are changes in a customer's circumstances that prevent them from dealing with their bank (e.g., financial conditions), and this represents a small percentage [9–11]; deliberate churn is caused by various factors, including new technological services, better prices, and quality factors [13].

Accordingly, banks should regularly monitor customers to detect warning signs regarding a customer's behavior that may lead to churn. Currently, researchers and bank mangers study patterns and trends in the data to develop models that can predict whether a customer is planning to churn or not [14]. Furthermore, data provide vital tools in banks; to discover the hidden patterns within large databases, it is recommended to apply clustering procedures, including neural network classification based on customer features. These procedures assist in building churn prediction models [15–18].

Customer churn prediction utilizing big data is a research area within machine learning technology, which works to classify distinctive types of customers into either churning or non-churning customers [14,19,20]. Many studies in the literature have created multiple prediction models relying on statistical and data mining techniques (machine learning models), such as linear regression, decision trees, random forests, logistic regression, neural networks, support vector machines, and deep neural networks [1,2,5,21,22].

The prediction of credit card customer churn is not a new field; many researchers have developed various prediction models. Kaya et al. (2018) [23] developed a prediction model which considered the individual transaction records of customers. The model mainly used information related to spatiotemporal factors, as well as choice and behavioral trait factors. The results showed that the developed model had more accurate prediction than the traditional models that considered demographics-based features. Moreover, early researchers tried to address the question of how machine learning models could be developed to predict customer churn, as discussed in Miao and Wang (2022) [24]. The authors developed a credit card customer churn prediction model by considering three machine learning approaches: random forest, linear regression, and k-nearest neighbor (KNN). The collected dataset contained 10,000 datum with 21 features, and the model was evaluated using the ROC, AUC, and confusion matrix. The results found that random forest had the best performance compared with the other machine learning models, with 96.3% accuracy. The results revealed that the top three important variables were the total transaction amount, the count in the last 12 months, and the total revolving balance. In addition, de Lima Lemos et al. (2022) [25] investigated churned customers in the banking sectors in Brazil. The study aimed to understand and predict the main variables that affected the customers who closed or stopped their accounts in the last six months. The study used various machine learning models, including random forest, k-nearest neighbor, decision tree, logistic regression, elastic net, and support vector machine models. The results found that random forest outperformed the other machine learning models in several performance metrics. The results found that customers with a stronger relationship with an institution, who borrowed more from the bank, were less likely to close their accounts. The results revealed that the model successfully predicted the losses of up to 10% of the operating results reported by the largest banks in Brazil.

Moreover, retail banking churn prediction was considered by Bharathi et al. (2020) [26]; the sample covered 602 young adult bank customers. To validate the developed model, different machine learning models were used, including ridge classifier cross-validation, the k-nearest neighbor classifier, decision tree, logistic regression, support vector classifier (SVC), and linear SVC. The results showed that the extra-tree classifier model had the highest performance compared with other models. The research showed that the top

features to develop a prediction model were the absence of mobile banking, zero-interest personal loans, zero balance, and other services online. Saias et al. (2022) [27] developed a churn risk prediction model for customers of cloud service providers. The aim of the study was to create an alert system to avoid losing cloud customers based on neural network model AdaBoost and random forest. The results found that random forest outperformed other models. Moreover, many researchers developed a prediction system for different churned customers in various fields such as the telecommunication industries [28], and the E-commerce industry [29].

Bank churn prediction aims to understand the possibility of customers moving from one bank to another. The reasons for movement include the availability of the latest technology, low interest rates, services offered, and credit card benefits [30]. This study aims to predict churned customers based on credit card and customer information (i.e., age and gender). Meanwhile, researchers have attempted to find credit card fraud detection using machine learning models [31,32] or by developing optimization algorithms with machine learning models [33].

Most of the prediction models in the literature have focused on developing prediction models for various problems in the banking system with a little interest in credit card customers. Therefore, the authors of this paper tried to fill the gaps of the previous studies in the field. The contributions of this paper are as follows:

1. Prediction models developed based on forwarding different numbers of independent variables.
2. The capability of the prediction models was validated based on two-step clustering and k-nearest neighbors.
3. The capabilities of the machine learning models to predict credit card customer churn in banks were predicted.
4. The top features for developing a credit card churn prediction method were determined.

The primary aims of this article are as follows:

1. To use different independent variables in building a prediction model based on two-step clustering and k-nearest neighbors.
2. To select the appropriate machine learning models with top features for predicting churn customers.

The rest of the paper is organized as follows: Section 1 presents the introduction and explores previous research in this field. Section 2 presents the research methodology used in this paper. Section 3 presents the analysis and empirical results. Finally, our conclusions are drawn in Section 4.

2. Research Methodology

Various studies have developed models for predicting customer churn without utilizing significant variables. To overcome this issue, it has been suggested that categorical variables are merged into one variable. Therefore, this research gap prompted the authors to find an appropriate model for predicting customer churn.

The primary step for developing a customer churn prediction model is to collect, analyze, and clean the dataset. Poorly cleaned data are unable to establish a relationship between input and output variables; in turn, this affects the performance of the prediction model. Therefore, the cleaned dataset can be applied in three models to build customer churn prediction models. The methods aim to select input variables depending on different independent variables that are selected by feeding all the independent variables in the dataset (continuous and categorical), selecting variables based on two-step clustering and logistic regression (continuous and cluster number variable), and selecting variables based on a feature-selection method. The outputs of the three models are applied in various machine learning models, including random forest, neural network, CR-Tree, C5 tree, Bayesian network, CHAID tree, support vector machine, quest tree, multinomial logistic regression, and a linear regression model. For brevity, only the top five machine learning models were

considered in this study. This section is divided into subsections: data collection, developed prediction models, machine learning models, and performance metrics.

2.1. Data Collection

This paper depends on the dataset of credit card customer churn for banks; the dataset was collected from https://leaps.analyttica.com (accessed on 10 September 2021). Customers have the option to choose one of four credit card types: blue, silver, gold, or platinum. When customers decide to change their bank, they are recorded as churn customers. Consequently, churn customers cause the profits of a banking system to decrease. Therefore, there has been increased interest from banking professionals to design an early-warning system to classify a bank's customers into churn or non-churn customers. The system would be able to notify the bank's managers so that they can communicate with customers who are expected to churn to improve their services, which is an appropriate way to keep the customers satisfied with their bank. The dataset contains 20 variables: 1 dependent variable and 19 independent variables. The total number of customers is 10,127, with 1627 churn customers.

The dataset contains the following data: a churn value (dependent), age, gender, number of dependents, education level, marital status, income category, product variable (type of credit card—blue, silver, gold, platinum), period of relationship with bank, total number of products held by the customer, number of months inactive of the last 12 months, number of contacts in the last 12 months, credit limit on the credit card, total revolving balance on the credit card, open to buy credit line (average of last 12 months), change in transaction amount (Q4 over Q1), total transaction amount (last 12 months), total transaction count (last 12 months), change in transaction count (Q4 over Q1), and average card utilization ratio. The full analysis of the dataset can be found in the DAS link.

Firstly, the dataset was divided into categorical and continuous variables as independent variables and one dependent variable (churn customers). Next, we analyzed the variables using different statistical metrics including min, max, variance, standard deviation, chi square (for categorical variable), and correlation analysis (for continuous variable). The initial analysis showed that the linear relationship between variables does not exist; therefore, a nonlinear model was applied to develop a prediction model for customer churn.

2.2. Developed Customers Churn Prediction Models

To develop a prediction model, three methods were utilized to control the number of independent variables used in the prediction model. First, all independent variables were directed to one of the applied machine learning models, which is referred to as Model 1. Next, we improved the independent variables by applying a two-step clustering method to the categorical variables only. Afterwards, the continuous variables with the cluster values were forwarded to each one of the machine learning models. To make the model more realistic, a logistic regression approach was used to predict the cluster number based on the categorical variables. This model is denoted as Model 2.

Model 2 was divided into two phases: the clustering phase and the prediction phase. In the clustering phase, the dataset was divided into groups using two-step clustering, and the continuous variables with a group variable were used to build a prediction model using the neural network. In the clustering model, the categorical variables were used to divide the customers into a specific number of groups. This step aimed to minimize the number of input variables forwarded to the machine learning model, in addition to simplifying the meta data of the customers. Moreover, the k-nearest neighbor model was used to ensure that the developed model was suitable for the online scenario and to avoid repeating the clustering analysis for future data. The k-nearest neighbors model uses customers' categorical meta data as input and the cluster number from the two-step clustering step as output. Furthermore, in the prediction phase, one of the machine learning models was used to build a prediction model by receiving the inputs from the two-step clustering

step and the continuous meta data of the customers. To build a prediction model which depends on machine learning, the dataset should be divided into training, validating, and testing data. The training data was used to learn the network from the previous information about the churn and non-churn customers. The test data were used to test the capability of the machine learning model in predicting the future churn customers. Based on previous research, the best percentages for building training, validating, and testing data for a dataset were found to be 70%, 15%, and 15%, respectively. Figure 1 shows how all the proposed phases of building the prediction model collaborated.

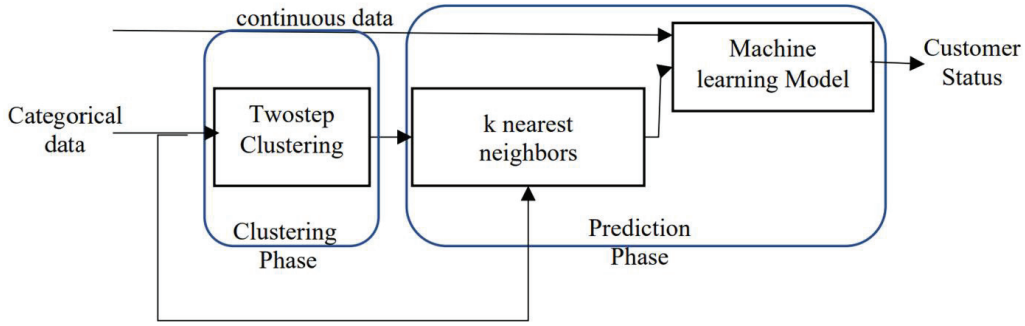


Figure 1. The proposed Model 2 based on k-nearest neighbors and two-step clustering.

Two-step clustering is a tool designed to handle the nature of the data and to find main insights. The differences between the two-step method and other clustering models include the following: it can use both categorical and continuous variables; then, it can automatically choose the appropriate number of clusters. Grouping data using the two-step clustering method involves an initial use of the distance measure to divide the data into groups; then, probabilistic approach is applied to select the optimal group.

For Model 3, all the independent variables, including the categorical and continuous variables, were forwarded to the feature-selection method to select the features that were related to the churn customers. The feature-selection method ranked the features as important, marginal, and unimportant. Only the important variables were forwarded to the machine learning models to be used in building the prediction models. A summary of the applied models is shown in Table 1.

Table 1. Summary of the developed models.

Model	Variables
Model 1	All variables
Model 2	All continuous variables and cluster value
Model 3	The selected variables after the feature-selection method

2.3. Machine Learning Models

This study adopted three methods for selecting independent variables, which aimed to understand the most suitable model for improving the performance of the prediction model based on machine learning models. The study applied ten machine learning models: random forest, neural network, CR-Tree, C5 tree, Bayesian network, CHAID tree, support vector machine, quest tree, multinomial logistic regression, and a linear regression model. The initial results denoted that the top five machine learning algorithms for the developed models in Table 1 were Bayesian network, C5 tree, CHAID tree, CR-Tree, and the neural network [34–38].

2.4. Performance Metrics

To design a predictor, the dataset was divided into training, validating, and testing datasets, with cutoff percentages of 70%, 15%, and 15% for training, testing, and validating the data, respectively.

The performance of churn customers prediction models can be evaluated by using classification parameter variables: *recall*, *precision*, *accuracy*, false omission rate, and *F1 score*. To find the performance metrics, a confusion matrix is generated first using the output of the classification results. The confusion matrix contains four parts: true positive (*TP*), true negative (*TN*), false positive (*FP*), and false negative (*FN*). The definitions of the confusion matrix are:

- True positives (*TP*): Number of churn customers correctly predicted as true churn.
- True negatives (*TN*): Number of non-churn customers correctly predicted as non-churn.
- False positives (*FP*): Number of churn customers incorrectly predicted as non-churn.
- False negatives (*FN*): Number of non-churn customers incorrectly predicted as churn.

To calculate performance metrics, the following equations are used:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$False\ Omission\ rate\ (FOR) = \frac{TN}{TN + FN} \tag{4}$$

$$F1\ score = 2 * \frac{Precision\ *\ Recall}{Precision + Recall} \tag{5}$$

3. Results Discussion and Analysis

To validate the developed models, the top five machine learning models were considered: Bayesian network, C5 tree, CHAID tree, CR-Tree, and the neural network. This section is divided based on the developed models explained previously in the methodology section.

3.1. Model 1: Develop Customers Churn Based on All Variables

Model 1 used all the categorical and continuous variables in the dataset. The independent variables were age, gender, number of dependents, education level, marital status, income category, product variable, period of relationship with bank, total number of products, number of months inactive in the last 12 months, no. of contacts in the last 12 months, credit limit on the credit card, total revolving balance on the credit card, open to buy credit line (average of last 12 months), change in transaction amount (Q4 over Q1), total transaction amount (last 12 months), total transaction count (last 12 months), change in transaction count (Q4 over Q1), and average card utilization ratio. The dependent variable was the status of the customer (churn or non-churn).

The training, testing, and validation results for Model 1 are shown in Table 2. The results showed that, for the training dataset, the accuracy and the FOR variables achieved more than 0.9, whereas the precision, recall, and F1 score showed variation between the machine learning models, with the best results achieved for the C5 tree model. To improve the performance of the training model, a validation dataset was used. The initial results showed that the Bayesian network, CHAID, and the neural network failed to enhance the performance of the prediction model. Additionally, the C5 tree and CR-Tree models showed improvements for the validation dataset compared with the training dataset, with best results derived using the C5 tree model. Finally, the dataset was tested to check the capability of the developed model to predict the future dataset. All the models' results show that their capabilities in predicting the churn customers is acceptable, with the highest performance derived with the C5 tree model. The accuracy, precision, recall, FOR, and F1 score for the C5 tree model with the testing dataset were 0.964, 0.914, 0.880, 0.974, and 0.897, respectively. The results indicated that the C5 tree model showed higher capability

in predicting the training, validating, and testing datasets compared with the other four machine learning models applied.

Table 2. The developed models based on all variables.

Dataset	Models	Accuracy	Precision	Recall	FOR	F1 Score
Train	Bayesian Network	0.935	0.831	0.740	0.952	0.783
Train	C5-Tree	0.966	0.918	0.860	0.974	0.888
Train	CHAID	0.906	0.776	0.569	0.923	0.656
Train	CR-Tree	0.924	0.750	0.785	0.959	0.767
Train	Neural Network	0.913	0.806	0.595	0.927	0.684
Test	Bayesian Network	0.941	0.865	0.790	0.955	0.826
Test	C5-Tree	0.964	0.914	0.880	0.974	0.897
Test	CHAID	0.910	0.824	0.629	0.924	0.713
Test	CR-Tree	0.929	0.773	0.854	0.968	0.811
Test	Neural Network	0.919	0.854	0.655	0.929	0.742
Validate	Bayesian Network	0.935	0.826	0.728	0.952	0.774
Validate	C5-Tree	0.975	0.961	0.873	0.977	0.915
Validate	CHAID	0.913	0.758	0.632	0.935	0.689
Validate	CR-Tree	0.930	0.750	0.816	0.966	0.782
Validate	Neural Network	0.905	0.784	0.526	0.919	0.630

To merge the collected results and the studied variables, the importance of the variables included as independent variables is shown in Figure 2. The importance of each variable analysis revealed that the top three variables were: change in transaction count, total revolving balance on the credit card, and total transaction count. Here, open-to-buy credit lines and the total transaction amounts showed very weak relations in building the C5 tree prediction model.

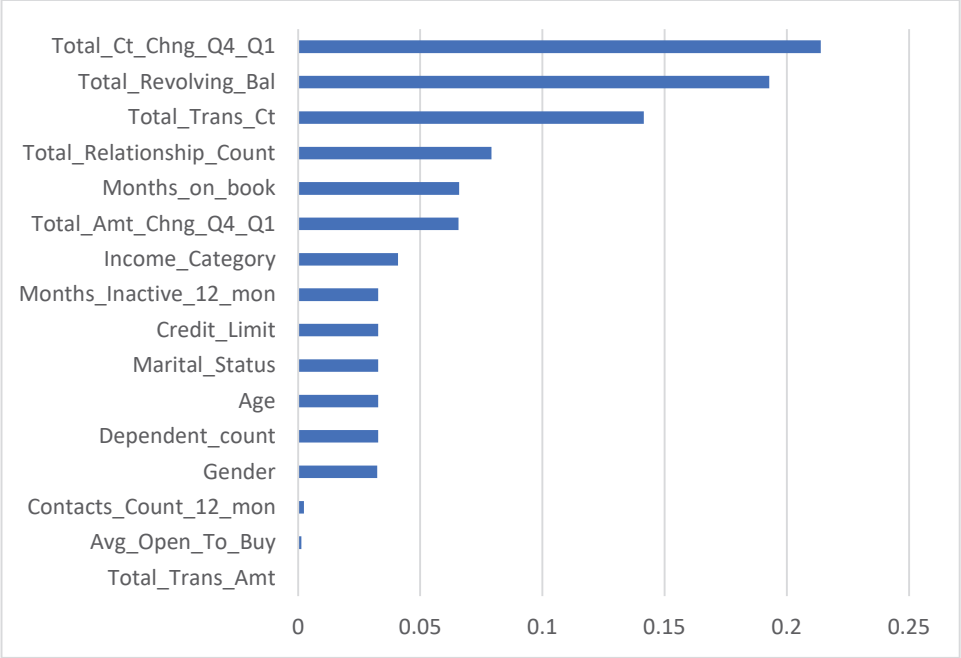


Figure 2. The importance of variables for Model 1 after applying C5 tree model.

3.2. Model 2: Developed Churn Customers Based on Two-Step Clustering and K-Nearest Neighbors

As discussed in the methodology section, the independent variables were divided into continuous and categorical variables. The categorical variables included gender, education level, marital status, income category, and product variable. The rest of the independent

variables were continuous. The categorical variables were forwarded to the two-step clustering model to merge the variables into one categorical variable that described all the categorical variables in the dataset. The results showed that the two-step clustering approach can create four categories for the input variables, as shown in Figure 3. The generated clusters are used with the k-nearest neighbor model to ensure that the developed model is similar to the real environment; here, k is the number of clusters generated using two-step clustering, as shown in the study methodology. The predicted values were used in the three datasets to validate the capability of the k-nearest neighbor model in predicting cluster values using the categorical variables. The accuracy results of the training, testing, and validating datasets were 0.93, 0.94, and 0.95, respectively. The results from the k-nearest neighbors approach indicated that the model is capable of predicting the generated cluster value using the categorical value.



Figure 3. Number of clusters after using two-step clustering.

The predicted values were forwarded with all the continuous values to one of the five machine learning models to develop a prediction model, as shown in Table 3.

Table 3. Two-step clustering and k-nearest neighbors for the developed model.

Dataset	Models	Accuracy	Precision	Recall	FOR	F1 Score
Train	Bayesian Network	0.931	0.817	0.726	0.949	0.769
Train	C5 Tree	0.963	0.911	0.851	0.972	0.880
Train	CHAID	0.906	0.776	0.569	0.923	0.656
Train	CR-Tree	0.924	0.750	0.785	0.959	0.767
Train	Neural Network	0.932	0.838	0.708	0.946	0.767
Test	Bayesian Network	0.941	0.868	0.787	0.955	0.825
Test	C5 Tree	0.967	0.919	0.891	0.977	0.905
Test	CHAID	0.910	0.824	0.629	0.924	0.713
Test	CR-Tree	0.929	0.773	0.854	0.968	0.811
Test	Neural Network	0.936	0.849	0.779	0.953	0.813
Validate	Bayesian Network	0.929	0.812	0.702	0.947	0.753
Validate	C5 Tree	0.975	0.953	0.882	0.979	0.916
Validate	CHAID	0.913	0.758	0.632	0.935	0.689
Validate	CR-Tree	0.930	0.750	0.816	0.966	0.782
Validate	Neural Network	0.923	0.796	0.667	0.941	0.726

The results from Model 2 show that the accuracy and FOR variables for all models were more than 0.9. The precision, recall, and F1 score variables were between 0.70 and 0.97 for the training, testing, and validation datasets. The training dataset showed that the C5 tree achieved the highest performance, where the accuracy, precision, recall, FOR, and F1 score were 0.963, 0.911, 0.851, 0.972, and 0.880, respectively. To improve the performance of the prediction model, a validation dataset was used. The validation results showed that the C5 tree and CR-Tree models improved the performance of the prediction models, while the rest of the prediction models did not. Moreover, to validate the performance of the future data, a testing dataset was used. The results showed that the C5 tree model had the highest performance compared with the other models. The test results of the C5 model were 0.967, 0.919, 0.891, 0.977, and 0.905 for accuracy, precision, recall, FOR, and F1 score, respectively. Finally, Model 2 showed that the C5 tree approach was more robust in creating a prediction model for churn customers.

To merge the collected results and the studied variables, the importance of each variable applied in the C5 tree model is shown in Figure 4. The results revealed that the top three variables in building the C5 tree prediction model were total transaction count, total revolving balance on the credit card, and change in transaction count; here, age- and month-inactive variables were very weak variables in building a C5 tree prediction model. In addition, the new cluster value generated in Model 2 was found to be among the top six most important variables in developing the customer churn prediction model.

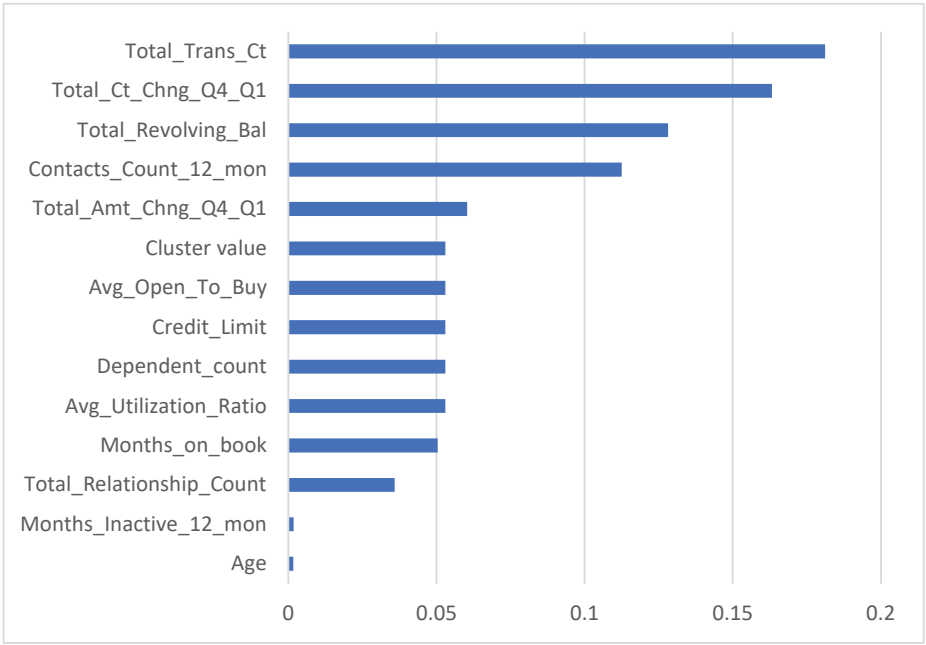


Figure 4. The importance of the variables for Model 2 after applying C5 tree model.

3.3. Model 3: Developed Churn Customers Based on Feature-Selection Model

Model 3 used a feature-selection model to select the most important variable(s) for the customers churn prediction (i.e., dependent variable). The output of the feature-selection model was divided into important, marginal, and unimportant, with cutoff percentages greater than 0.95, 0.90, and 0, respectively, as shown in Table 4. The feature-selection model used a Person model for the categorical variable.

The results showed that Total_Trans_Ct, Total_Ct_Chng_Q4_Q1, Total_Revolving_Bal, Contacts_Count_12_mon, Avg_Utilization_Ratio, Total_Trans_Amt, Total_Relationship_Count, Months_Inactive_12_mon, and Total_Amt_Chng_Q4_Q1 are the important variables to be used in building Model 3. The rest of the variables were omitted, as they were either marginal or unimportant. The important variables were forwarded to five machine learning models to develop the customer churn prediction model.

Table 4. The output of the feature-selection method.

Status	Rank	Field	Importance	Value	Importance
TRUE	1	Total_Trans_Ct	0	1	important
TRUE	2	Total_Ct_Chng_Q4_Q1	0	1	important
TRUE	3	Total_Revolving_Bal	0	1	important
TRUE	4	Contacts_Count_12_mon	0	1	important
TRUE	5	Avg_Utilization_Ratio	0	1	important
TRUE	6	Total_Trans_Amt	0	1	important
TRUE	7	Total_Relationship_Count	0	1	important
TRUE	8	Months_Inactive_12_mon	0	1	important
TRUE	9	Total_Amt_Chng_Q4_Q1	0	1	important
FALSE	10	Gender	1	0.95	Marginal
FALSE	11	Age	1	0.93	Marginal
FALSE	12	Dependent_count	2	0.85	Unimportant
FALSE	13	Income_Category	2	0.83	Unimportant
FALSE	14	Education_Level	2	0.76	Unimportant
FALSE	15	Avg_Open_To_Buy	2	0.76	Unimportant
FALSE	16	Months_on_book	2	0.65	Unimportant
FALSE	17	Credit_Limit	2	0.52	Unimportant
FALSE	18	Marital_Status	2	0.26	Unimportant
FALSE	-	Card_Category			Removed

The results of Model 3 showed that all five models predicted the churned customers with different accuracy values, ranging between 0.9 and 0.99 for all the collected datasets, as shown in Table 5. The training dataset shows that the C5 tree model was the most accurate model in building a prediction model for customer churn. The training results were 0.976, 0.921, 0.928, 0.986, and 0.924 for accuracy, precision, recall, FOR, and F1 score, respectively. To improve the performance of prediction models, a validation dataset was applied. The validation results showed that the trained models were not optimized, indicating that the validation dataset was not able to improve the performance of the trained models. To check the predictability of the machine learning model, the testing dataset was used. The test results showed that all five models could accurately predict the churn customers with different performance metrics. The best results were supported by applying the C5 tree prediction model, where the accuracy, precision, recall, FOR, and F1 score results were 0.940, 0.813, 0.861, 0.970, and 0.836, respectively.

Table 5. The developed customer churn model based on feature selection.

	Models	Accuracy	Precision	Recall	FOR	F1_Score
Train	Bayesian Network	0.929	0.823	0.701	0.945	0.757
Train	C5 Tree	0.976	0.921	0.928	0.986	0.924
Train	CHAID	0.905	0.779	0.559	0.921	0.651
Train	CR-Tree	0.929	0.785	0.763	0.956	0.774
Train	Neural Network	0.925	0.830	0.667	0.939	0.739
Test	Bayesian Network	0.930	0.855	0.730	0.943	0.788
Test	C5 Tree	0.940	0.813	0.861	0.970	0.836
Test	CHAID	0.905	0.838	0.581	0.915	0.686
Test	CR-Tree	0.937	0.816	0.831	0.963	0.824
Test	Neural Network	0.926	0.839	0.723	0.942	0.777
Validate	Bayesian Network	0.926	0.821	0.662	0.941	0.733
Validate	C5 Tree	0.944	0.814	0.825	0.968	0.819
Validate	CHAID	0.902	0.752	0.544	0.921	0.631
Validate	CR-Tree	0.927	0.763	0.763	0.957	0.763
Validate	Neural Network	0.917	0.813	0.592	0.930	0.685

To investigate the optimal variables used in the C5 tree customers churn prediction model, Figure 5 is used to explain the importance of each variable applied in the C5 tree model. The results showed that total transaction count, the total revolving balance on the credit card, and the change in the transaction count are the most important variables for predicting churn customers; the total transaction amount is not considered in the customer churn prediction model.

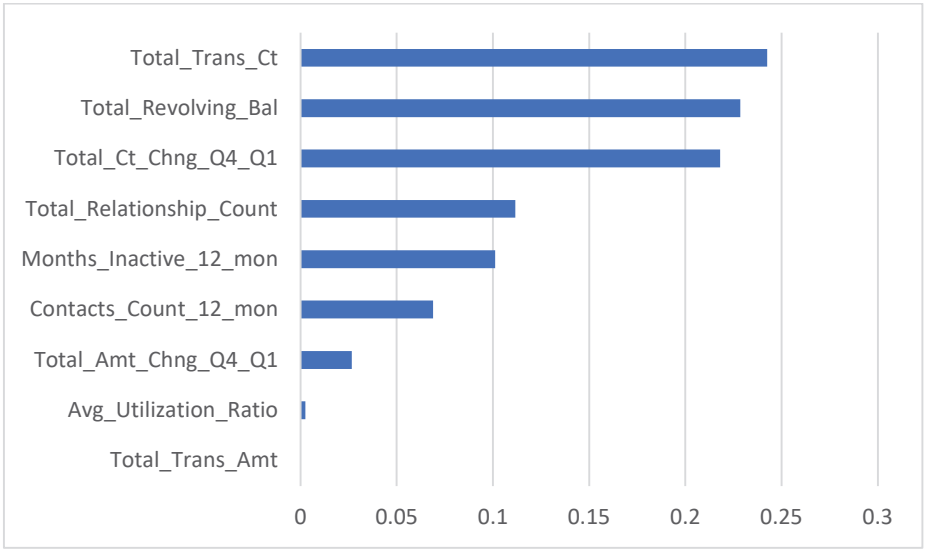


Figure 5. The importance of variables for Model 3 after applying C5 tree customer churn model.

3.4. Discussion and Analysis

Three models were adopted to develop a machine learning model that can be used to predict churn customers. The models were built after considering the independent variables selection method. To select the variables, three models were considered: all variables selection (Model 1), two-step clustering and k-nearest neighbor selection (Model 2), and a feature-selection algorithm (Model 3). The results revealed that all the applied machine learning models were capable of predicting the churn customers with different efficiency values. In all the developed models (Models 1–3), the C5 tree prediction model showed the highest performance in customer churn prediction, with the best performance derived by Model 2. The accuracy, precision, recall, FOR, and F1 score using the training dataset were 0.967, 0.919, 0.891, 0.977, and 0.905, respectively. These results indicated that the C5 tree model is highly capable of predicting the churn customers, with fewer variables than Model 1 and extra analysis compared with Model 3. The results showed that the top three variables for all the developed models in building the C5 tree model were the total transaction count, the total revolving balance on the credit card, and the change in the transaction count. In addition, the results showed that the total transaction amount was not important in developing a prediction model.

However, achieving the highest performance is not always the goal; thus, researchers can obtain good performance with a less complex model by using a feature-selection model with the given dataset. Accordingly, they can reduce the number of independent variables that are forwarded to the machine learning models. In addition, using all the variables with the C5 tree model can guarantee good performance with more processing time to predict the churn customers. Overall, the results indicated that the machine learning models have a high capability to adapt to different numbers of input variables in predicting the churn customers. In addition, the C5 tree model showed a higher capability of predicting the churn customers. Finally, determining the appropriate number of independent variables is a very important step in developing and building a machine learning model.

All models can be used to predict credit card customer churn in banking systems. Model 1 requires that all independent variables are used to predict the churn cases. Model 2

requires all continuous variables and one combined categorical variable to predict the churn cases. Model 3 uses only variables with high importance in predicting the churn cases. The models clearly indicate the desired direction which an expert should take to retain the bank's clients. Experts and scholars in the field can use all the models to predict churn customers, and Model 1 can be used if nonexperts use the model, given that all variables will be included in the model. Models 2 and 3 can be used by experts in the field because of the additional analyses required in developing prediction models for customer churn.

4. Conclusions

This paper aimed to investigate the capability of machine learning to predict the credit card customer churn rate in the banking sector. The collected dataset contains two types of data: categorical and continuous variables. These data describe 10,127 customers in the bank; 1627 of these customers were churn customers. To develop different credit card customer churn prediction models, the independent variables were changed to different forms. To change the number of independent variables, three models were proposed. The models were named Model 1 (all variables—categorical and continuous variables), Model 2 (two-step clustering and k-nearest neighbors), and Model 3 (feature-selection model). In addition, five machine learning models were suggested: Bayesian network, C5 tree, CHAID, CR-Tree, and a neural network. Then, the original dataset was divided into three datasets: training, testing, and validation, comprising 70%, 15%, and 15% of the original dataset, respectively. The three models were used with five machine learning models and the results showed that the machine learning models were capable of predicting the credit card customer churn. The results supported that, for Models 1–3, the C5 tree model outperformed all other machine learning models. The results revealed that the total transaction count, the total revolving balance on the credit card, and the change in the transaction count are the top three important variables to develop any churn customer prediction model. In addition, it is not necessary to use all the categorical variables (such as gender, education level, marital status, income category, and product variable) in developing a churn customer prediction model. On the other hand, adding a single categorical variable describing all the categorical variables in the dataset can improve the performance of the churn customer prediction model. The results also revealed that, to improve the churn customer prediction model, the selection of independent variables is an important step. This step can be implemented using one of the feature-selection models, or a combination of several variables. The credit card customer churning rate is the percentage of a bank's customers who try to leave the service. Building an early-prediction model to predict the status of a bank's customers may help them to avoid losing their customers.

In summary, the results indicated that clustering with KNN and the C5 tree model outperformed previous models in various performance metrics, including R^2, precision, and recall. Clustering the independent variables can improve the prediction performance in various sciences [39–42], and the number of transactions is dominant in identifying churn customers. The results proved that C5 tree models can outperform other functional models [16,41,42].

In future work, analyses to further understand the optimal independent variables are required to develop a more robust, more accurate, faster, less complicated, and more efficient churn prediction model. Such a study will cover more datasets with extra variables to extract the most important variables; in addition, new machine learning models will be executed to determine which is the optimal model. The use of the conventional machine learning model will not always guarantee the best results for a given dataset. Therefore, the accuracy and efficiency of prediction models should be improved in the future.

Author Contributions: Formal analysis, H.A.-N.; investigation, D.A.-N., N.A.-R. and H.A.-N.; methodology, D.A.-N., N.A.-R. and H.A.-N.; writing—original draft preparation, D.A.-N., N.A.-R. and H.A.-N.; writing—review and editing, D.A.-N., N.A.-R. and H.A.-N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found here: [from https://leaps.analyttica.com (accessed on 10 September 2021)].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jagadeesan, A.P. Bank customer retention prediction and customer ranking based on deep neural networks. *Int. J. Sci. Dev. Res.* **2020**, *5*, 444–449.
2. Amuda, K.A.; Adeyemo, A.B. Customers churn prediction in financial institution using artificial neural network. *arXiv* **2019**, arXiv:1912.11346.
3. Kim, S.; Shin, K.-S.; Park, K. An application of support vector machines for customer churn analysis: Credit card case. In Proceedings of the International Conference on Natural Computation, Changsha, China, 27–29 August 2005; Springer: Berlin/Heidelberg, Germany; pp. 636–647.
4. Kumar, D.A.; Ravi, V. Predicting credit card customer churn in banks using data mining. *Int. J. Data Anal. Tech. Strateg.* **2008**, *1*, 4–28. [CrossRef]
5. Keramati, A.; Ghaneei, H.; Mirmohammadi, S.M. Developing a prediction model for customer churn from electronic banking services using data mining. *Financ. Innov.* **2016**, *2*, 10. [CrossRef]
6. Bastan, M.; Akbarpour, S.; Ahmadvand, A. Business dynamics of iranian commercial banks. In Proceedings of the 34th International Conference of the System Dynamics Society, Delft, The Netherlands, 17–21 July 2016.
7. Bastan, M.; Bagheri Mazrae, M.; Ahmadvand, A. Dynamics of banking soundness based on CAMELS Rating system. In Proceedings of the 34th International Conference of the System Dynamics Society, Delft, The Netherlands, 17–21 July 2016.
8. Iranmanesh, S.H.; Hamid, M.; Bastan, M.; Hamed Shakouri, G.; Nasiri, M.M. Customer churn prediction using artificial neural network: An analytical CRM application. In Proceedings of the International Conference on Industrial Engineering and Operations Management, Bangkok, Thailand, 5–7 March 2019; pp. 23–26.
9. Domingos, E.; Ojeme, B.; Daramola, O. Experimental analysis of hyperparameters for deep learning-based churn prediction in the banking sector. *Computation* **2021**, *9*, 34. [CrossRef]
10. Chen, S.C.; Huang, M.Y. Constructing credit auditing and control & management model with data mining technique. *Expert Syst. Appl.* **2011**, *38*, 5359–5365.
11. Hadden, J.; Tiwari, A.; Roy, R.; Ruta, D. Computer assisted customer churn management: State-of-the-art and future trends. *Comput. Oper. Res.* **2007**, *34*, 2902–2917. [CrossRef]
12. Risselada, H.; Verhoef, P.C.; Bijmolt, T.H. Staying power of churn prediction models. *J. Interact. Mark.* **2010**, *24*, 198–208. [CrossRef]
13. Kim, H.S.; Yoon, C.H. Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. *Telecommun. Policy* **2004**, *28*, 751–765. [CrossRef]
14. Xia, G.; He, Q. The research of online shopping customer churn prediction based on integrated learning. In Proceedings of the 2018 International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2018), Qingdao, China, 30–31 March 2018; pp. 30–31.
15. Olaniyi, A.S.; Olaolu, A.M.; Jimada-Ojuolape, B.; Kayode, S.Y. Customer churn prediction in banking industry using K-means and support vector machine algorithms. *Int. J. Multidiscip. Sci. Adv. Technol.* **2020**, *1*, 48–54.
16. Nie, G.; Rowe, W.; Zhang, L.; Tian, Y.; Shi, Y. Credit card churn forecasting by logistic regression and decision tree. *Expert Syst. Appl.* **2011**, *38*, 15273–15285. [CrossRef]
17. Seng, J.L.; Chen, T.C. An analytic approach to select data mining for business decision. *Expert Syst. Appl.* **2010**, *37*, 8042–8057. [CrossRef]
18. Tsai, C.F.; Lu, Y.H. Customer churn prediction by hybrid neural networks. *Expert Syst. Appl.* **2009**, *36*, 12547–12553. [CrossRef]
19. Rahman, M.; Kumar, V. Machine learning based customer churn prediction in banking. In Proceedings of the 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 5–7 November 2020; IEEE: Piscataway, NJ, USA; pp. 1196–1201.
20. Khodabandehlou, S.; Rahman, M.Z. Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behaviour. *J. Syst. Inf. Technol.* **2017**, *19*, 65–93. [CrossRef]
21. Miguéis, V.L.; Van den Poel, D.; Camanho, A.S.; e Cunha, J.F. Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert Syst. Appl.* **2012**, *39*, 11250–11256. [CrossRef]
22. Kolajo, T.; Adeyemo, A.B. Data Mining technique for predicting telecommunications industry customer churn using both descriptive and predictive algorithms. *Comput. Inf. Syst. Dev. Inform. J.* **2012**, *3*, 27–34.
23. Kaya, E.; Dong, X.; Suhara, Y.; Balcisoy, S.; Bozkaya, B. Behavioral attributes and financial churn prediction. *EPJ Data Sci.* **2018**, *7*, 41. [CrossRef]
24. Miao, X.; Wang, H. Customer churn prediction on credit card services using random forest method. In Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022), Online, 14–16 January 2022; Atlantis Press: Paris, France; pp. 649–656.

25. de Lima Lemos, R.A.; Silva, T.C.; Tabak, B.M. Propension to customer churn in a financial institution: A machine learning approach. *Neural Comput. Appl.* **2022**, *4*, 11751–11768. [CrossRef]
26. Bharathi, S.V.; Pramod, D.; Raman, R. An ensemble model for predicting retail banking churn in the youth segment of customers. *Data* **2022**, *7*, 61. [CrossRef]
27. Saias, J.; Rato, L.; Gonçalves, T. An approach to churn prediction for cloud services recommendation and user retention. *Information* **2022**, *13*, 227. [CrossRef]
28. Thakkar, H.K.; Desai, A.; Ghosh, S.; Singh, P.; Sharma, G. Clairvoyant: AdaBoost with cost-enabled cost-sensitive classifier for customer churn prediction. *Comput. Intell. Neurosci.* **2022**, *2022*, 9028580. [CrossRef] [PubMed]
29. Xiahou, X.; Harada, Y. Customer churn prediction using AdaBoost classifier and BP neural network techniques in the E-commerce industry. *Am. J. Ind. Bus. Manag.* **2022**, *12*, 277–293. [CrossRef]
30. Nie, G.; Wang, G.; Zhang, P.; Tian, Y.; Shi, Y. Finding the hidden pattern of credit card holder's churn: A case of china. In Proceedings of the International Conference on Computational Science, Vancouver, BC, Canada, 29–31 August 2009; Springer: Berlin/Heidelberg, Germany; pp. 561–569.
31. Kulatilleke, G.K. Challenges and complexities in machine learning based credit card fraud detection. *arXiv* **2022**, arXiv:2208.10943.
32. Alfaiz, N.S.; Fati, S.M. Enhanced credit card fraud detection model using machine learning. *Electronics* **2022**, *11*, 662. [CrossRef]
33. Jovanovic, D.; Antonijevic, M.; Stankovic, M.; Zivkovic, M.; Tanaskovic, M.; Bacanin, N. Tuning machine learning models using a group search firefly algorithm for credit card fraud detection. *Mathematics* **2022**, *10*, 2272. [CrossRef]
34. Al-Najjar, D.; Assous, H.F.; Al-Najjar, H.; Al-Rousan, N. Ramadan effect and indices movement estimation: A case study from eight Arab countries. *J. Islam. Mark.* **2022**. *ahead-of-print*. [CrossRef]
35. Al-Najjar, D.; Al-Najjar, H.; Al-Rousan, N. Evaluation of the prediction of COVID-19 recovered and unrecovered cases using symptoms and patient's meta data based on support vector machine, neural network, CHAID and QUEST Models. *Eur. Rev. Med. Pharmacol. Sci.* **2021**, *25*, 5556–5560.
36. Al-Rousan, N.; Al-Najjar, H.; Alomari, O. Assessment of predicting hourly global solar radiation in Jordan based on Rules, Trees, Meta, Lazy and Function prediction methods. *Sustain. Energy Technol. Assess.* **2021**, *44*, 100923. [CrossRef]
37. Al-Najjar, D.; Al-Najjar, H.; Al-Rousan, N.; Assous, H.F. Developing Machine Learning Techniques to Investigate the Impact of Air Quality Indices on Tadawul Exchange Index. *Complexity* **2022**, *2022*, 1–12. [CrossRef]
38. Al-Najjar, H.; Al-Rousan, N. A classifier prediction model to predict the status of Coronavirus COVID-19 patients in South Korea. *Eur. Rev. Med. Pharmacol. Sci.* **2020**, *24*, 3400–3403.
39. Rajamohamed, R.; Manokaran, J. Improved credit card churn prediction based on rough clustering and supervised learning techniques. *Clust. Comput.* **2018**, *21*, 65–77. [CrossRef]
40. AL-Rousan, N.; Mat Isa, N.A.; Mat Desa, M.K.; AL-Najjar, H. Integration of logistic regression and multilayer perceptron for intelligent single and dual axis solar tracking systems. *Int. J. Intell. Syst.* **2021**, *36*, 5605–5669. [CrossRef]
41. Al-Najjar, H.; Alhady, S.S.N.; Saleh, J.M. Improving a run time job prediction model for distributed computing based on two level predictions. In Proceedings of the 10th International Conference on Robotics, Vision, Signal Processing and Power Applications, Pulau Pinang, Malaysia, 14–15 August 2018; Springer: Singapore; pp. 35–41.
42. Al-Najjar, H.; Alhady, S.S.N.; Mohamad-Saleh, J.; Al-Rousan, N. Scheduling of workflow jobs based on twostep clustering and lowest job weight. *Concurr. Comput. Pract. Exp.* **2021**, *33*, e6336. [CrossRef]

sustainability

MDPI

Article

Digital Inclusive Finance and Family Wealth: Evidence from LightGBM Approach

Ying Liu [1,2,3], Haoran Zhao [1], Jieguang Sun [2,*] and Yahui Tang [2]

[1] School of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun 130117, China
[2] Key Laboratory of Financial Technology of Jilin Province, Changchun 130117, China
[3] Business Big Data Research Center of Jilin Province, Changchun 130117, China
* Correspondence: 401065@jlufe.edu.cn

Abstract: With the rapid development of digital technology in China, Digital Inclusive Finance, which uses digital financial services to promote financial inclusion, is developing rapidly. This paper uses the Peking University Digital Financial Inclusion index of China and China Family Panel Studies (CFPS) data to construct a predictive model using the LightGBM machine learning algorithm to study whether Digital Inclusive Finance can predict household wealth and analyze the characteristics of strong predictive ability for household wealth. They found that: (1) the introduction of the Digital Financial Inclusion index can improve the prediction performance of the household wealth model; (2) financial literacy and age characteristics are the key characteristics of household wealth accumulation; (3) the coverage and depth of Digital Inclusive Finance has a significant effect on family wealth accumulation, but the degree of digitization acts as a disincentive factor. This paper not only uses machine learning methods to do research on Digital Inclusive Finance and family wealth from a more comprehensive perspective, but also provides effective theoretical support for the key factors that enhance family wealth.

Keywords: family wealth; machine learning; Digital Inclusive Finance

1. Introduction

In recent years, the popularity of Digital Inclusive Finance has largely improved the efficiency of financial services, as well as promoted innovation, consumption, and national income, which has had a huge impact on socioeconomic effects. China established inclusive finance as a national strategy for the first time after the State Council promulgated the Plan for Promoting the Development of Inclusive Finance (2016–2020). Inclusive finance refers to the provision of appropriate and effective financial services at affordable costs for all segments and groups of society in need. After the approval of the G20 Advanced Principles for Digital Inclusive Finance at the G20 Hangzhou Summit in 2016, China began to focus on promoting the integration of inclusive finance and digital technology, specifically financial products and services such as payments, credit, insurance, securities, and savings, through digital tools for transactions. Since then, China has entered the era of digital economy. Digital Inclusive Finance adapts to the scenario of traditional financial inclusive finance where there is little information, scattered distribution, small customer size, and lack of collateral. Digital Inclusive Finance has social and economic effects, such as improving financial service efficiency, optimizing resource allocation, and promoting innovation, consumption, and income growth, because it can solve the problems of information asymmetry, high cost, and high risk [1]. China has made remarkable achievements in economic and social development so far, and the national family wealth and comprehensive national capacity have increased significantly.

The family is an important demand side in the financial market, and it is an important element of social aggregation. Family wealth, as the material basis for a better life for the

people, plays an important role in maintaining the stability of social and economic structure and alleviating the imbalance of development [2–4]. Therefore, it is important to study the effects of the impact of Digital Inclusive Finance on family wealth accumulation and to meet the demand for financial accessibility of families in all social classes and regions.

At present, most of the current literature on family wealth and Digital Inclusive Finance uses explanatory methods, which use statistical models to infer the causal relationship between variables based on hypothesis testing. These methods manually assume a specific functional form of the model beforehand, such as exponential relationships, linear relationships, and so on. However, many models that fit well have no specific functional form set in advance. In contrast, predictive models using machine learning algorithms start from the data itself, and gradually train the model to fit the data best by iterating many times according to the principle of constantly finding the optimal solution, so that more complex relationships between variables can be uncovered and patterns between the data can be explained, thus helping to advance the relevant theories. The machine learning model itself belongs to the black box model, but scholars improve the interpretability of machine learning algorithms by studying alternative variables, which improves the prediction effect of machine learning, and also supplements the theoretical category of interpretative methods [5].

In view of this, this paper designs a household wealth prediction indicator system, based on a large number of papers related to Digital Inclusive Finance and household wealth, using 38,459 household panel data and 77,697 individual panel data from 31 provinces and municipalities directly under the central government of China from 2014 to 2018. Furthermore, the LightGBM machine learning algorithm is used to verify whether Digital Inclusive Finance can predict household wealth. To analyze the relationship with household wealth in multiple dimensions to improve the model interpretation, the Accumulated Local Effects Plot algorithm is proposed to further explore the main factors affecting family wealth accumulation in three dimensions: breadth of coverage, depth of use, and degree of digitization. The remainder of the paper is organized as follows: (1) a review of the literature related to Digital Inclusive Finance, family wealth, and machine learning algorithms applied to the fields of finance and management, as well as providing a theoretical basis for the design of model indicator systems in Section 2; (2) introduce the data sources and the specific indicator system constructed in Section 3; (3) show the specific algorithm used, the experimental procedure, and the experimental results in Section 4; (4) analysis and discussion for experimental results in Section 5; and (5) finally, conclusions.

2. Review of the Literature

2.1. Factors Affecting Family Wealth

Regarding the analysis of factors affecting family wealth accumulation, scholars have conducted more adequate research, mainly including two perspectives of family characteristics and family member characteristics. In studying the household asset portfolio in China, Lei and Zhou [6] found that the health factors of urban residents have an important impact on guiding the distribution of family wealth. This was later confirmed in the findings of Xie [7], while this study found a significant effect of education and knowledge literacy, and whether or not self-employment had a differential impact on household wealth. However, the process in the study regarding the impact of knowledge literacy on household wealth has not been shown. Wu et al. [4] also argue that knowledge literacy can promote household wealth accumulation, as shown by the fact that households with higher financial literacy allocate more assets to financial assets, especially risky financial assets. Furthermore, this study also argues that household size has different degrees of positive effects on household wealth accumulation. There are also many examples in the literature that analyze this characteristic of household location and confirm that urban–rural differences in households and geographical differences lead to different asset allocation structures of households, which have an impact on household wealth [8–10] In the literature that has been studied from the perspective of family member's characteristics, some scholars have found that

individual cognition [11,12] and individual education [13,14] allow individuals to understand financial instruments more comprehensively and allocate assets more rationally. It has also been found that mental health [15] individual self-control [16], childhood family structure [17], and gender [18] all have important effects on the accumulation of household wealth. From the above studies, it is clear that previous studies on household wealth have focused on certain characteristics to explore their relationship with household wealth, and there are few studies that have analyzed them under a comprehensive framework. In addition, the previous literature mostly measures and studies households and family members separately, while the impact indicator systems of households and family members are different, so there is a need to analyze them together. In this paper, based on the above-mentioned literature, we conduct the construction of a specific indicator system for the impact factors of household wealth, so as to prepare the premise for the experiments of the modeling method afterwards.

2.2. Digital Inclusive Finance and Family Wealth

Dedicated to providing financial services to groups excluded from the financial system, financial inclusion has grown to some extent worldwide. There is a range of literature that demonstrates the positive effects of financial inclusion on promoting social status equity [19], access to banking services [20], alleviating distributional inequalities [21], and optimizing social behavior outcomes [22]. Along with the advancement of digital technology, Digital Inclusive Finance has emerged, and its development breaks through time and space constraints and reduces the dependence of traditional transaction models on physical space. Currently, the Digital Inclusive Finance development index has risen from 40 in 2011 to 341.22 in 2020. Digital Financial Inclusion has enlivened the economy and quickly penetrated all aspects of household life [23,24], influencing household economic behavior and becoming a new factor affecting household wealth growth [25]. More and more scholars are exploring the impact of digital inclusion on countries (regions) from a macro perspective, for example, by looking at rural areas as an entry point. Song empirically analyzed the income gap between urban and rural residents in 31 Chinese provinces using the Theil index, and found that Digital Inclusive Finance can significantly reduce the urban–rural income gap [26]. Using multidimensional poverty as an entry perspective, Zhang et al. demonstrated that financial inclusion improves the viability of rural families, and thereby improves the quality of poverty reduction by reducing the likelihood of future poverty among rural households [27]. Based on the perspective of urban–rural differences, Zhang and Wu argued that Digital Financial Inclusion can narrow the gap between urban and rural incomes, improve the transmission mechanism of income distribution, and stimulate the entrepreneurial dynamics of rural residents, especially for small and micro or labor-intensive entrepreneurial activities [28]. Moreover, some scholars have also studied the impact of Digital Inclusive Finance on income from the perspectives of entrepreneurship [29], farmers' income increase [30], distribution patterns [31], and human capital [32]. In addition to macro analysis, some studies focused on micro households to explore the impact of Digital Inclusive Finance development on household financial asset allocation [4], entrepreneurship [33–35], residential consumption [27], household risk [36], credit [37], and other effects. As well as the positive effects of digital inclusion, some scholars argue that digital inclusion can have an impact on exacerbating the gap between rich and poor. Hu found that the lack of digital tools for low-income households made it difficult to participate in the process of Digital Inclusive Finance, which, in turn, suppressed low-income household income and increased relative poverty [38]. Direct studies use household net worth as the explanatory variable and discuss less analysis of factors of Digital Inclusive Finance for family wealth growth [3]. The above studies use explanatory models to test the causal relationship between variables and prove the positive effect of Digital Inclusive Finance development on promoting family wealth accumulation; more commonly used are structural equation and linear regression models [25,34,39,40]. On the one hand, explanatory models require specific functional assumptions and restrictions on

the relationship between data and variables, but most real-world problems have difficulty assuming functional form among variables [41]; on the other hand, explanatory models are prone to overfitting in order to promote the perfect fitting of sample data when dealing with high-dimensional data, resulting in their poor extrapolation prediction [42] In the context of Digital Inclusive Finance, the characteristics that affect household wealth are complex and multidimensional, and the characteristics themselves can be relational. This is difficult to describe using traditional linear fitting models. Therefore, it is necessary to explore new and effective methods to deeply explore the factors of family wealth accumulation.

2.3. The Application of Machine Learning on Family Wealth Issues

Machine learning is an important area of artificial intelligence that stems from statistical model fitting. Machine learning uses inference and sample learning to derive appropriate theory from data, especially for solving problems such as "noisy" patterns and large data sets. It plays an increasingly important role in the analysis of large samples, multi-vectors, and uncertain data [43,44]. The main features of machine learning methods include: (1) machine learning can get the results that fit the data best by continuously learning and looping to optimize the target problem; (2) the fitted model can help us explore more statistical relationships between data system characteristics and variables, and can find more complex patterns in the data [45]; (3) machine learning uses a range of methods, such as regularization and pruning, to better solve the overfitting problem [46], which is difficult to achieve with explanatory models. Some scholars have now used machine learning methods to deal with certain economic-type problems, such as stock return prediction [47,48], the effect of executive characteristics on firm performance [49,50], commercial bank liquidity monitoring [51], the issue of changing organizational boundaries of firms and the selection of collaborative innovation partners [52,53], etc. However, there is less research literature on the analysis of factors influencing family wealth accumulation using machine learning methods.

Combing the above literature, we find that, firstly, there is a large amount of literature examining the impact of Digital Inclusive Finance on asset allocation and narrowing the wealth gap, but less literature examining the impact of Digital Inclusive Finance on household income and household wealth. Secondly, the studies focus more on the problem itself, few focus on the modeling methods, and lack of systematic and quantitative findings from the predictive capability. Under this background, this paper studies the impact of Digital Inclusive Finance on household wealth accumulation based on the LightGBM model. The main contribution of this study is that it can better identify the regular pattern and main factors that affect the accumulation of family wealth. At the same time, the proposed model also enriches the measurement methods of family wealth models.

3. Description of Data Sources and Variables

3.1. Data Source

The first part of the data is the "Digital Inclusive Finance Index of Peking University", which is jointly compiled by Peking University Digital Finance Research Center and Ant Group Research Institute, and the data set contains 33 specific indicators in 3 dimensions. The second part of the data is the China Family Panel Studies (CFPS) of the China Social Science Survey Center of Peking University, which has been conducted every two years since 2010, and the data covers three levels of individuals, families, and communities. Based on the need of model setting and data availability, individual and household data of three periods in 2014, 2016, and 2018 were selected, extracted, and combined, and finally, a total of 77,697 sample data were obtained for the three periods.

3.2. Variable Definition

From the existing literature, the definition of household wealth is unclear. In this paper, we refer to Hurst and Lusardi [54] Gai et al. [55] (2013), Gan et al. [56], and other

scholars' methods of defining household wealth and the characteristics of the database. Finally, family net worth is used as a proxy variable for household wealth.

Digital Inclusive Finance consists of three dimensions: the breadth of coverage of digital finance, the depth of use of digital finance, and the degree of digitization of inclusive finance. The breadth of coverage of digital finance is expressed in the provision of adequate digital financial services. The depth of use of digital finance refers to the effective demand for digital financial services. The degree of digitization of inclusive finance is the convenience, low cost, and creditability of financial services. If Digital Inclusive Finance can contribute to household wealth growth, then breadth of reach, depth of use, and digitalization can also contribute to household wealth growth.

We filtered the rest of the core variables in this paper on the basis of the literature in Section 2 and summarized them into two main categories. One is the family-level characteristics, which include family size, urban–rural classification, whether working in agriculture, whether self-employed, and the amounts of books in the family collection. The other is the characteristics of family member-level, which can be divided into objective characteristics and subjective characteristics according to the ease of observation. Objective characteristics include: age, marital status, gender, health status, education level, whether accessing the Internet, and other variables that can be easily measured objectively. Subjective characteristics include: satisfaction with life, satisfaction with the environment, perception of the wealth gap, perception of employment, perception of medical care, perception of education, perception of housing, perception of social security, and other variables that are difficult to measure directly.

3.3. Descriptive Statistics

The descriptive statistics of family-level characteristics and family member-level characteristics are shown in Table 1, which includes the sample size, mean, and standard deviation of each characteristic. From the results of the description of family-level characteristics: the average net assets of household are CNY 586,323.3; the average size of family is 3–4 persons; 50.24% of respondents live in urban areas; 48.06% of families are engaged in agricultural work; 9.40% of families are engaged in the self-employed private sector; the average number of books in the household collection is 68; and the average financial inclusion index of each region is 217.68. From the results of the description of family member-level characteristics, in terms of objective characteristics: the age span is disparate, the highest is 102 years old, the lowest is 16 years old, and the average age is 46; the percentage of men was 49.27 and the percentage of women was 50.73, thus the gender ratio is almost equal; the average education is between junior high school and high school; the percentage of respondents using the Internet was 40.93; 79.19% of respondents are married; and the average health level of the respondents was relatively healthy. In terms of subjective characteristics: the average view of respondents on various social issues is more positive, and views on life satisfaction, environmental issues, wealth gap, employment issues, medical issues, education issues, housing issues, and social security issues are all above the average level and more positive and optimistic.

Table 1. Variable description statistics.

Variable	Variable Definition	Sample Size	Average Value	Standard Deviation
	family-level			
NetAsset	Family net worth	38,459	586,323.3	1,570,973
FamilySize	Family size	38,459	3.647209	1.869107
AgBagd	Whether work in agriculture (Yes = 1; No = 0)	38,459	0.4806417	0.4996316
InPrivate	Whether the family is engaged in self-employed private (Yes = 1; No = 0)	38,459	0.0940482	0.2918996
InCity	Whether the household is located in a town (yes = 1; no = 0)	38,459	0.5024312	0.5000006
BookSum	Household book collection	38,459	67.38932	333.8275
Index*	Digital Financial Inclusion index	38,459	217.6776	52.54776
coverage_breadth*	Breadth of coverage index	38,459	188.4211	56.74687
usage_depth*	Depth of use index	38,459	215.0571	68.72808
digitization_level*	Degree of digitization index	38,459	319.0469	65.27698
	Family member-level			
NetAsset	Family net worth	77,697	597,015.2	1,526,159
Age	Age of family member	77,697	46.28097	16.47327
Gender	Gender (male = 1; female = 0)	77,697	0.4927089	0.4999501
InNet	Internet access or not (Yes = 1; No = 0)	77,697	0.409308	0.4917093
Marry	Whether married	77,697	1.46868	1.112772
Health	Health status (0–4)	77,697	1.997889	1.227294
Education	Academic qualifications (No need to study = 0; elementary school = 1; middle school = 2; high school = 3; college = 4; high school = 5; bachelor's degree = 6; master's degree = 7; doctorate = 8)	77,697	2.626742	1.50079
LifePoint	Satisfaction with life (1–5)	77,697	3.810919	1.027157
EvPoint	Satisfaction with the environment (0–10)	77,697	6.694274	2.672565
PgPoint	Views on the gap between the rich and the poor (0–10)	77,697	7.005058	2.402166
EmpPoint	Opinions on employment (0–10)	77,697	6.469658	2.4133
EduPoint	Views on education (0–10)	77,697	6.247938	2.681608
MedcPoint	Opinions on medical care (0–10)	77,697	6.317966	2.618888
HousPoint	Opinions on housing (0–10)	77,697	6.121897	2.710084
SocisePoint	Opinions on social security (0–10)	77,697	5.959805	2.636506
Index*	Digital Inclusive Finance index	77,697	212.9693	52.87561
coverage_breadth*	Breadth of coverage index	77,697	183.3167	57.07364
usage_depth*	Depth of use index	77,697	210.6925	67.90457
digitization_level*	Degree of digitization index	77,697	341.9367	66.21179

4. Research Methods and Model Construction

4.1. Research Methods

4.1.1. LightGBM

LightGBM (Light Gradient Boosting Machine) is an open source framework proposed by Microsoft in 2017 that combines the ideas of integrated learning and decision trees, and it is an improved gradient boosting algorithm based on the GBDT (Gradient Boosting Decision Tree) algorithm. GBDT is an iterative decision tree algorithm where each learning is based on the residuals of the previous training. LightGBM adopts the Histogram-based Algorithm, Gradient-based One-Side Sampling (GOSS), and Exclusive Feature Bundling (EFB) technology to reduce the running time of the algorithm complexity. When LightGBM processes high-latitude and massive data, it occupies less memory, and has higher computing speed and prediction accuracy [57]. The specific modeling process is as follows: firstly, the initial value of the model is set to 0, as in Equation (1):

$$\rho_0 = 0 \quad (1)$$

ρ_0 denotes the initial value, followed by the calculation of the optimal split point among all features according to the Histogram-based Algorithm, at this point, the first tree is obtained, as in Equation (2):

$$\rho_1 = \rho_0 + \psi_1 \tag{2}$$

ρ_1 represents the prediction model of the first round, and ψ_1 denotes the first tree. Based on the first tree, we calculate the tree that minimizes the objective function, which is called the second tree. Put the second tree behind the first tree and become the prediction model for the second round. The above steps are called an iterative process. Repeat the above iteration until the last tree, as in Equation (3):

$$\rho_0 = 0\rho_1 = \rho_0 + \psi_1\rho_2 = \rho_1 + \psi_2 \ldots \rho_i = \rho_{i-1} + \psi_i \tag{3}$$

4.1.2. Accumulated Local Effects Plot (ALE)

Interpretability is summarized as the degree to which humans can understand the reasons for model decisions [58] In current research, most machine learning methods tend to increase the complexity of the model in order to increase its own prediction accuracy, leading to a decrease in the interpretability of the model [59]. In this paper, we try to use the Accumulated Local Effects Plot algorithm [60] to deeply analyze the explanatory relationship between Digital Financial Inclusion and family wealth on the LightGBM model.

The ALE algorithm describes the impact mode of a single feature variable on the prediction results, that is, controlling other features as observed values, and modifying the target features to change the model fitting results. The core of the observed impact model is calculating the prediction bias. First, the value of the target feature is divided into multiple intervals according to quantiles to form a grid. Then, the effects of all target features in an interval on a single instance are summed, and the average effect in the interval is calculated at the same time. Finally, accumulating the average effect of all grid intervals to find the impact result. Explain Equation (4) as:

$$\hat{\tilde{f}}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{\mathrm{i}:x_j^{(i)} \epsilon N_j(k)} \left[f\left(z_{k,j}, x_{\setminus j}^{(i)}\right) - f(z_{k,j}, x_{\setminus j}^{(i)}) \right] \tag{4}$$

where, x is the target feature, j is the number of features, k is the number of intervals divided, $n_j(k)$ is the total amount of data in the kth interval, $Z_{k,j}$ is the upper and lower bound taken for the kth interval of the jth feature, and $i : x_j^{(i)} \epsilon N_j(k)$ is at the ith sample point of the kth interval. Centralize the effects, that is, let each effect subtract the average effect. When the feature effect is positive, indicating that the feature has a positive effect on the predicted outcome, the expression is shown in Equation (5):

$$\tilde{f}_{j,ALE}(x) = \hat{\tilde{f}}_{j,ALE}(x) - \frac{1}{n} \sum_{\mathrm{i}=1}^{n} \hat{\tilde{f}}_{j,ALE}\left(x_j^{(i)}\right) \tag{5}$$

4.2. Model Construction

4.2.1. Indicator System

In this paper, the family net worth (NetAsset) is set as the predicted variable, and the model is defined as a family model (Fa_Md) and a family member model (Fn_Md) according to the family characteristics and family member characteristics, and these two models are collectively called the Basic Model. In addition, the Digital Inclusive Finance index and its sub-indices are introduced in the Basic Model, including: breadth of coverage index, depth of use index, and degree of digitization index. The model, after the introduction of the Digital Inclusive Financial Index and its sub-indices, is defined as the Family Digital Inclusive Model (FaDi_Md) and the Family Member Digital Inclusive Model (FnDi_Md),

and the two models are collectively called the Digital Inclusive Model. The specific model composition, predictor variables, and predicted variable composition are shown in Table 2.

Table 2. Model indicator system.

Predicted Variables	Model		Predictive Variables
NetAssets	Basic Model	Family Model (Fa_Md)	FamilySize, AgBagd, InPrivate, InCity, BookSum
		Family Member Model (Fn_Md)	Age, Gender, InNet, Marry, Health, Education, LifePoint, EvPoint, PgPoint, EmpPoint, EduPoints, MedcPoint, HousPoint, SocisePoint
	Digital Inclusive Model	Family Digital Inclusive Model (FaDi_Md)	FamilySize, AgBagd, InPrivate, InCity, BookSum, Index*(coverage_breadth*, usage_depth*, digitization_level*)
		Family Member Digital Inclusive Model (FnDi_Md)	Age, Gender, InNet, Marry, Health, Education, LifePoint, EvPoint, PgPoint, EmpPoint, EduPoint, MedcPoint, HousPoint, SocisePoint, Index* (coverage_breadth*, usage_depth*, digitization_level*)

Note: The Digital Inclusion Index* is synthesized from the coverage_breadth*, the usage_depth*, and the digitization_level*.

4.2.2. Algorithm Implementation

Step 1 randomly splits the data set of each period of the family model, with part of the samples as the training data set and part of the samples as the test data set. Step 2 constructs the LightGBM model, and the whole process of the experiment is done by using single-period data training and next-period data rolling fit. In the learning process, we should adjust the learning rate, the number of iterations, the depth of the single-base regression tree, and other parameters of the model. Step 3 calculates the goodness of fit R^2 and the mean absolute error MAE to test the model performance. Step 4 calculates the ranking index using the relative importance built into LightGBM to obtain the relative importance features that affect family wealth accumulation. Step 5 introduces the cumulative local effect algorithm to further analyze the impact of the Digital Financial Inclusion sub-index on family wealth prediction and marginal effect. The specific algorithm implementation steps are shown below.

Step 1: Data splitting

The data set is split proportionally according to the experimental requirements to form a training data set and a test data set.

Step 2: Build the LightGBM model

① Build the model

Using the model indicator system in Table 2 and following the process of Equations (1)–(3) to construct four models, respectively, as follows:

$$\rho_I = \sum_{I=1}^{I} \psi_i(X, \theta_i) \tag{6}$$

where, ρ_I is the established LightGBM model. ψ_i is the ith tree, I is the total number of trees, X is the predictor variable dataset, and θ_i is the set of learning parameters for the ith tree.

② Objective function

Since the problem in this paper belongs to the regression problem, the model internal loss function chosen is MSE: $T(\rho_{obs}, \rho_i) = (\rho_{obs} - \rho_i)^2$, and the objective function of the model can be obtained as:

$$\Gamma_i = T(\rho_{obs}, \rho_i) + \Omega(f_i) \tag{7}$$

The objective function Γ_i contains two components, the first term is the loss function T, in which ρ_{obs} are the control observations; the second term Ω is the regularization term.

③ Regular items

The regular items are used to control the complexity of the model to avoid overfitting problems, and two penalty functions are added to the regular items. The specific canonical terms are as follows.

$$\Omega(f_i) = \lambda \sum_{j=1}^{J} |w_j| + \frac{1}{2}\alpha \sum_{j=1}^{J} w_j^2 \tag{8}$$

In which w_j is the number of leaf scores of the model split, w_j^2 is the L2 parametric number of leaf node scores, λ, α are coefficients, and the specific values can be adjusted in the process of use, and the default values are used in this paper.

④ Parameter adjustment

According to the experimental comparison results, the learning rate is set to 0.1, the number of iterations is 6000, and the depth of a single regression tree is 7.

⑤ Rolling verification

A single period training data set is used for training and prediction, and then the next period prediction data set is used for validation, rolling backwards in sequence by year for fitting.

Step 3: Evaluation metrics

Evaluation metrics: one is the goodness of fit R^2, and the other is the mean absolute error MAE. The goodness of fit R^2 is calculated as:

$$R^2 = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n}(y_i - \overline{y})^2} = 1 - \frac{RSS}{TSS} \tag{9}$$

The mean absolute error MAE is calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{10}$$

y_i is the true value of the test set, $\overline{y}$ and $\hat{y}_i$ are the mean value of the previous sample and the prediction value trained using the previous sample, respectively. RSS and TSS are the sum of squared residuals and total squared deviations between samples. R^2 takes the value range [0, 1], and the model's prediction ability is better when R^2 is closer to 1. MAE takes the value range [0, +∞], and the smaller the MAE, the better the prediction ability of the model.

Step 4: Relative importance ranking

In the splitting process of the decision tree in Equations (1)–(3), the individual feature weights and the number of times selected scores are counted, where the larger the feature is to the splitting point improvement performance measure (closer to the root node), the larger the weight is, the more decision trees are selected and the higher the score is. Finally, the results of a feature in all decision trees are averaged after weighted summation to obtain the importance score, and all features are ranked and processed accordingly.

Step 5: Accumulated local effects plot

The Digital Inclusive Finance index, the breadth of coverage index, the depth of use index, and the degree of digitization index are brought into Equations (4) and (5) as target features to measure the specific impact effects of the respective features with respect to the predicted features.

4.3. Empirical Research

4.3.1. A Comparison of Approaches to Digital Inclusive Finance for Household Wealth Building

The objectives of the experiments in this section include: (1) comparing the performance of machine learning methods and comparing the proposed model with the ordinary least squares (OLS) method; (2) verifying whether the introduction of the Digital Financial Inclusion index into the index system can improve the performance of household wealth prediction. The experimental data years span from 2014 to 2018, and the experiments are fitted in a rolling fashion, with the following specific experimental procedures: firstly, the data set is divided into two phases, 2014–2016 and 2016–2018; each year of data is split according to 7.5:2.5, with the first 75% data set training the model and the last 25% data set validating the model. For example, the model was trained using 75% of the 2014 dataset, 25% of the 2016 dataset was used to validate the model, and then the model was trained again using 75% of the 2016 dataset to predict the next period, and so on. The experimental results are shown in Table 3, and it is easy to find that the LightGBM method outperforms the OLS method for both the Base Model (Fa_Md, Fn_Md) and the Digital Inclusive Model (FaDi_Md, FnDi_Md). Taking the Family Member Digital Inclusive Model (FnDi_Md) as an example, for the 2014–2016 and 2016–2018 data for both periods, the R^2 fitted using the LightGBM method is 20.72% and 19.96%, respectively, while the R^2 results of the OLS method are 11.83% and 14.10%, which are 8.89% and 5.86% higher, respectively. Similarly, for the MAE evaluation index, LightGBM obtains 27.82% and 36.08%, while the OLS results are 39.02% and 39.76%, which are 11.2% and 3.68% lower, respectively. The above analysis shows that the LightGBM method performs better than the traditional statistical model in fitting the data, so it is feasible to use the LightGBM machine learning method to explore the relationship between digital financial inclusion and household wealth.

Table 3. Fitting results.

Evaluation Methods	Crossover Year	Evaluation Indicators	Model Fitting Results				Degree of Promotion	
			Basic Model		Digital Inclusive Model			
			Fa_Md	Fn_Md	FaDi_Md	FnDi_Md	FaDi_Md–Fa_Md	FnDi_Md–Fn_Md
OLS	14–16	R^2 (%)	1.82	3.51	4.33	11.83	2.51	8.32
		MAE (%)	37.40	38.37	39.09	39.02	1.69	0.65
	16–18	R^2 (%)	7.01	1.76	18.19	14.10	11.18	12.34
		MAE (%)	39.16	45.13	35.73	39.76	−3.43	−5.37
LightGBM	14–16	R^2 (%)	4.33	5.82	11.36	20.72	7.03	14.9
		MAE (%)	31.15	34.19	26.92	27.82	−4.23	−6.37
	16–18	R^2 (%)	8.49	3.83	27.34	19.96	18.85	16.13
		MAE (%)	38.41	42.34	31.81	36.08	−6.6	−6.26

This part analyzes whether the Digital Inclusion Model can improve the predictive power of the model for household wealth to a greater extent, using the Base Model as a control group. According to the fitting results of LightGBM in Table 3, it can be observed that for the two periods of data from 2014 to 2016 and 2016 to 2018, the R^2 of the Family Digital Inclusive Model (FaDi_Md) is improved by 7.03% and 18.85%, respectively, compared to the Family Model (Fa_Md) at the family level. Similarly, at the family member level, the Family Member Digital Inclusive Model (FnDi_Md) compared to the Family Member Model (Fn_Md) increased R^2 by 14.9% and 16.13%. The results of the Digital Inclusion Model outperformed the Base Model both at the family level and at the family member level, and the results suggest that the introduction of the Digital Financial Inclusion index can significantly improve the predictive power of family wealth.

4.3.2. Analysis of the Key Features of Digital Financial Inclusion for Family Wealth Growth

The experiments in the previous section compare and analyze the fitting effects of the Base Model and the Digital Inclusion Model, verifying that the Digital Financial Inclusion

index can improve the predictive power of household wealth, but the extent of its influential role in the prediction of results is not known. The purpose of the experiments in this section is to compare the relative importance of features in the Family Digital Inclusive Model (FaDi_Md) and the Family Member Digital Inclusive Model (FnDi_Md). Using the proposed model step 4 method, the scores obtained from feature selection during regression tree splitting were weighted and averaged and ranked, and the results are shown in Table 4. The experimental results found that for the Family Digital Inclusive Model (FaDi_Md), the family book collection (BookSum, 31.45%), Digital Financial Inclusion index (Index*, 24.95%), and family size (FamilySize, 24.20%) were ranked highly, while for the Family Member Digital Inclusive Model (FnDi_Md), the feature age (Age, 24.92%), education (Education, 21.89%), and Digital Financial Inclusion index (Index*, 10.05%) were more important. The results show that the Digital Financial Inclusion index (Index*) occupies an important position in both models because the technology platform of digital finance has a long-tail effect and low marginal cost, which can alleviate the information asymmetry phenomenon, reduce investment risk, enhance family investment participation, and thus promote the increase of household income and wealth growth. The family book collection (BookSum) reflects the overall knowledge and cultural literacy of the family, and the characteristic education is the education level of individual family members, which indicates that individual knowledge literacy and education level have a great role in promoting family wealth management, and family members with a higher literacy level are more conducive to the creation and increase of family wealth. The characteristic family size (FamilySize) indicates the number of family members. FamilySize indicates the number of family members who create wealth in the family. The age variable reflects the social experience, wealth accumulation experience, and wealth cognitive differences of individual family members. It means that the more people who create family wealth, the more conducive to family wealth accumulation. The elderly, with rich social experience as the leaders of family capital decision making, pay more attention to wealth accumulation.

Table 4. Feature importance order.

	Family Digital Inclusive Model (FaDi_Md)		**Family Member Digital Inclusive Model (FnDi_Md)**	
	Variables	**Feature Importance (%)**	**Variables**	**Feature Importance (%)**
1	BookSum	31.45	Age	24.92
2	Index*	24.95	Education	21.89
3	FamilySize	24.20	Index*	10.05
4	InCity	7.35	LifePoint	5.93
5	AgBagd	6.58	HousPoint	5.41
6	InPrivate	5.47	Health	5.34
7			EvPoint	3.99
8			InNet	3.98
9			MedcPoint	3.70
10			EmpPoint	3.53
11			EduPoint	3.13
12			PgPoint	2.64
13			Marry	2.55
14			SocisePoint	2.52
15			Gender	0.42

4.3.3. Multidimensional Analysis of the Effect of Digital Inclusive Finance on Family Wealth Growth

The above two experiments confirm that Digital Inclusive Finance is an important factor influencing household wealth accumulation from the perspective of algorithm and relative importance of features, respectively. Since the Digital Financial Inclusion index is synthesized from three sub-indices: breadth of coverage, depth of use, and degree of digitization, where breadth of coverage is reflected by sufficient digital financial products and financial services, depth of use is the effective demand for digital financial services,

and degree of digitization is expressed as the flexibility, low cost, and creditability of financial services [3]. This section of the experiment uses the ALE to further explore the explanatory relationship between digital financial inclusion and household wealth in different dimensions.

The ALE algorithm acts on the household digital inclusion model (FaDi_Md) and the household member digital inclusion model (FnDi_Md), respectively, to produce a graph of the Digital Financial Inclusion index and its sub-indices in relation to household wealth (as shown in Figures 1–4). Figures 1a–c and 2a–c are the analysis graphs of the Family Digital Inclusive Model and the Family Member Digital Inclusion Model analysis diagram for 2014–2016. Figures 3a–c and 4a–c are the analysis diagram of the Family Digital Inclusive Model and the Family Member Digital inclusive Model for 2016–2018, respectively. The results of the four graphs show that as the Digital Financial Inclusion index, depth of use, and breadth of coverage increase, household wealth shows an increasing trend. The relationship between the Digital Financial Inclusion index, depth of use, breadth of coverage, and household wealth shows a more significant positive correlation. The results indicate that the development of Digital Inclusive Finance is conducive to increase the overall wealth of households, and that the breadth of coverage and depth of use elements of Digital Inclusive Finance are the main reasons for the positive correlation in the results. This may be due to the fact that digital inclusion offers a wide range of financial products and services through information technology, which lowers the threshold for households to participate in financial activities, thereby boosting household economic dynamics and increasing the total value of household wealth.

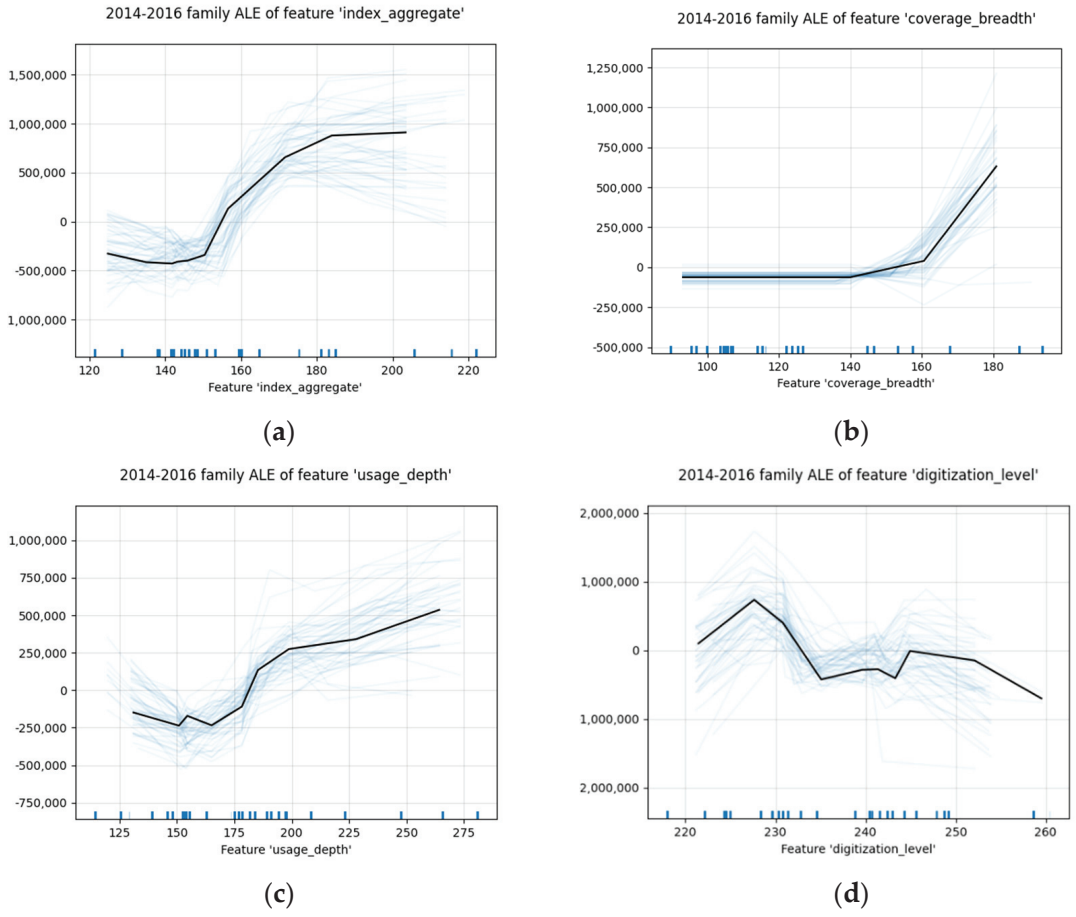


Figure 1. ALE analysis of Family Digital Inclusive Model (2014–2016). (**a**–**d**) represent the relationship between digital inclusion index, breadth of coverage index, depth of use index, and digitalization index and household wealth, respectively.

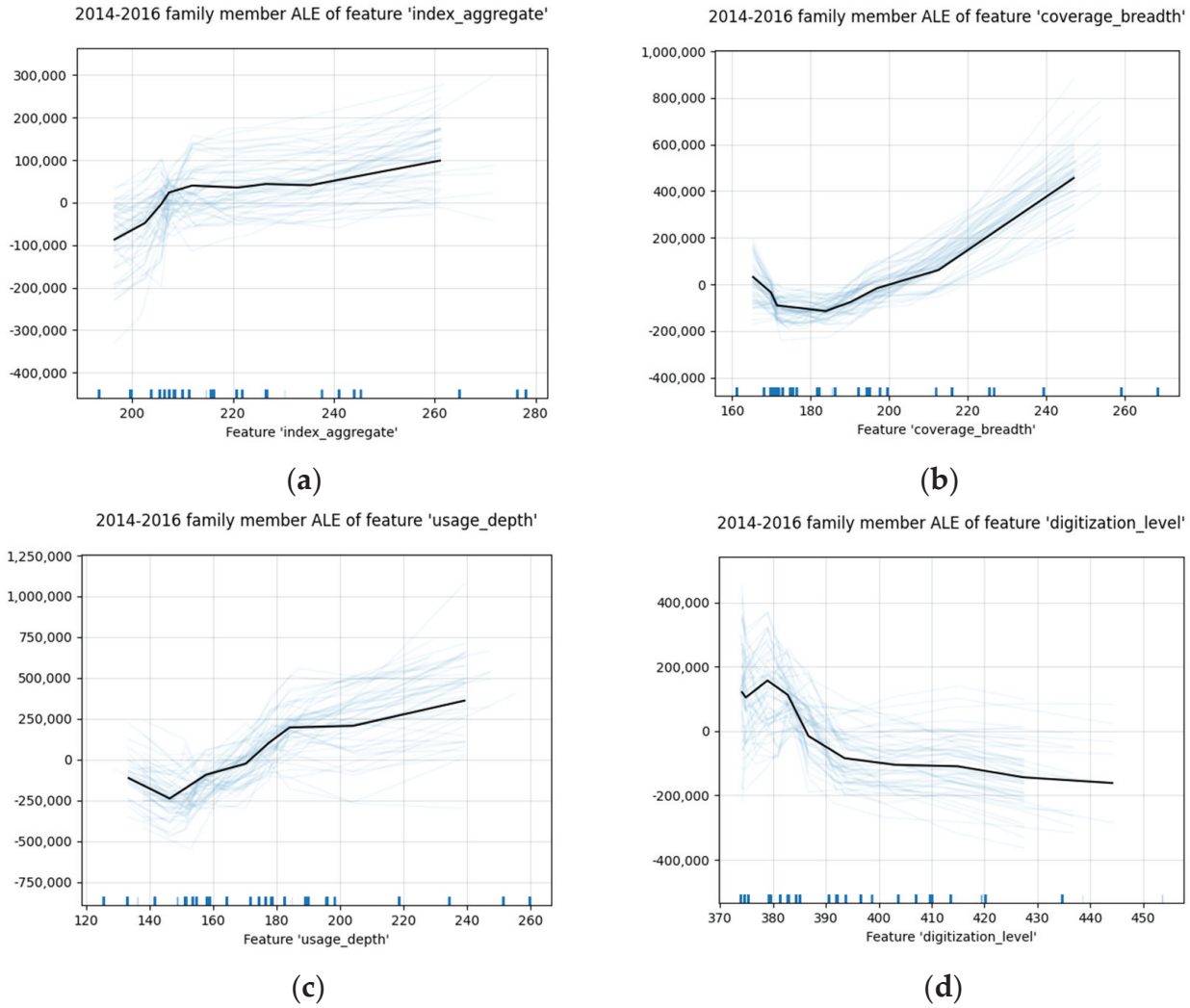


Figure 2. ALE analysis of Family Member Digital Inclusive Model (2014–2016). (**a**–**d**) represent the relationship between digital inclusion index, breadth of coverage index, depth of use index, and digitalization index and household wealth, respectively.

On the other hand, Figures 1d and 2d show that there is no obvious linear relationship between the degree of digitization and household wealth, indicating that the degree of digitization has less effect on family wealth and even has a suppressive effect within a certain range, while Figures 3d and 4d show the negative correlation between the degree of digitization and household wealth. This phenomenon indicates that there is a "digital divide" in Digital Inclusive Finance. The main reasons: (1) digital financial services rely excessively on Internet connection, which makes them unable to provide services to people who do not have digital devices and have low access to digital technology; (2) due to the different education levels of households, there is a cognitive bias towards the products and services provided by financial institutions, which leads to financial exclusion; (3) network data vulnerability and network security problems also reduce the trust of customers in digital financial platforms. As a result of the "digital divide", the spatial environment for information access, participation and information sharing is hindered, and some groups are unable to successfully participate in economic and social activities, lacking opportunities and being isolated, thus aggravating the polarization between rich and poor and hindering the growth of household wealth.

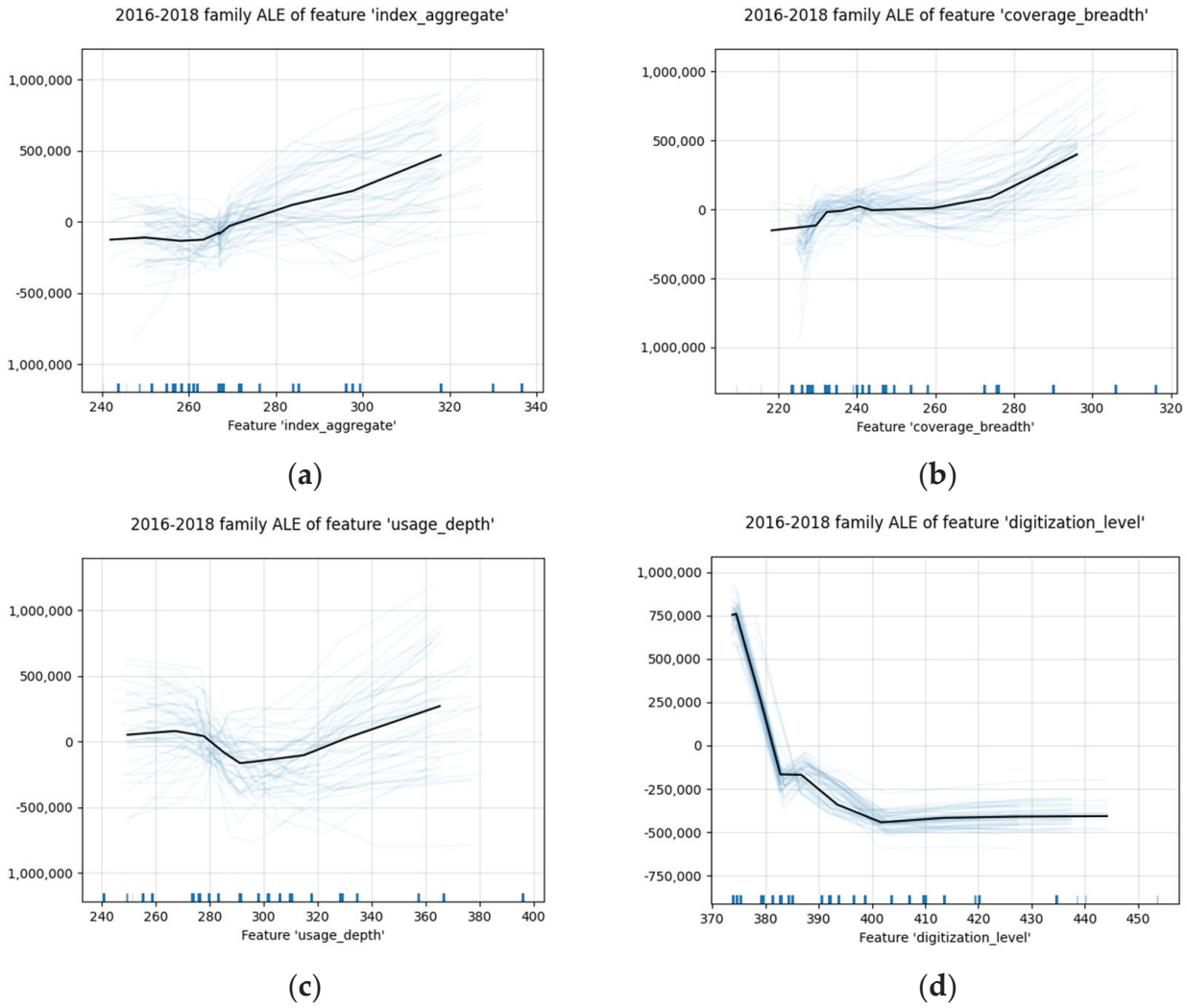


Figure 3. ALE analysis of Family Digital Inclusive Model (2016–2018). (**a**–**d**) represent the relationship between digital inclusion index, breadth of coverage index, depth of use index, and digitalization index and household wealth, respectively.

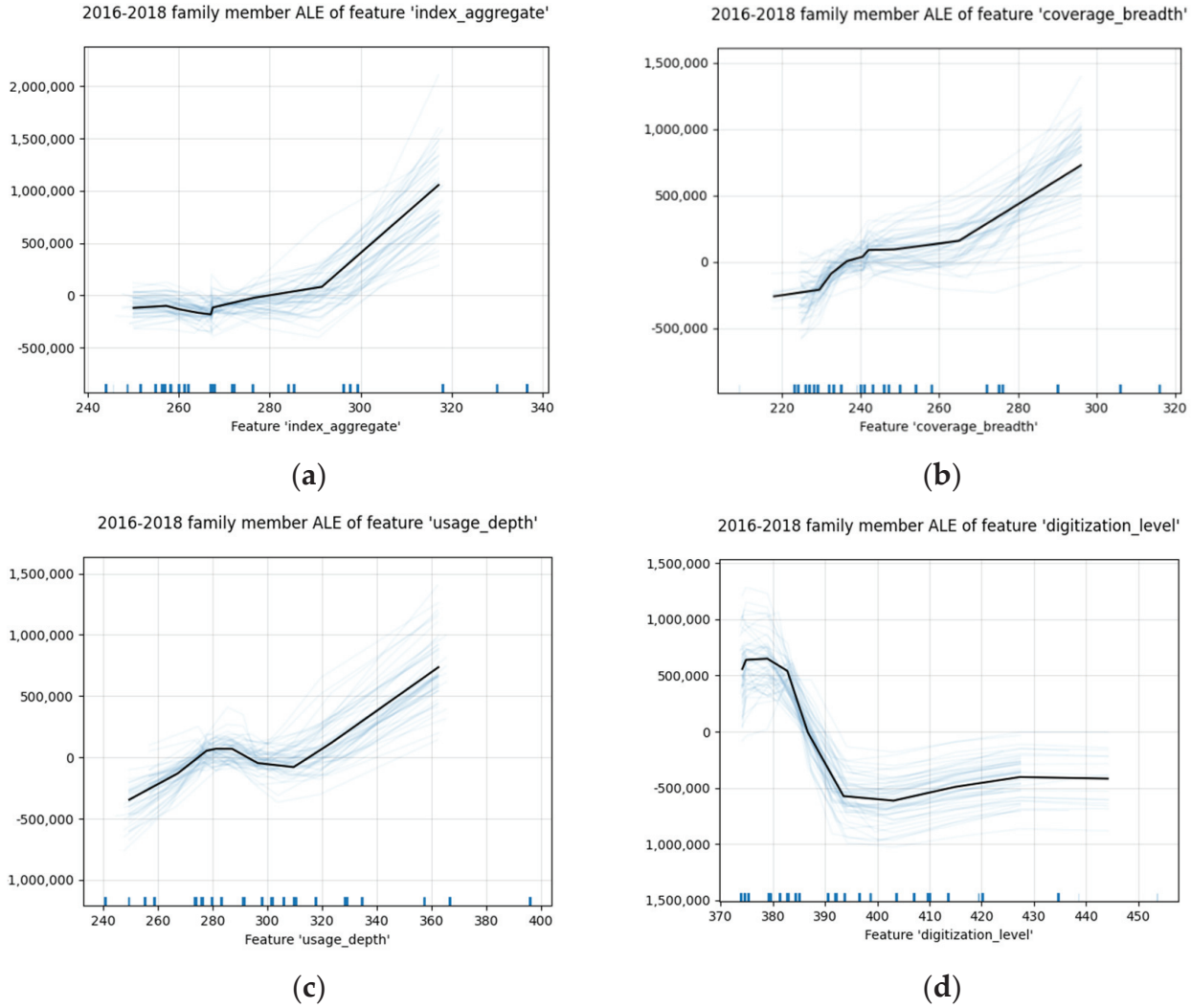


Figure 4. ALE analysis of Family Member Digital Inclusive Model (2016–2018). (**a**–**d**) represent the relationship between digital inclusion index, breadth of coverage index, depth of use index, and digitalization index and household wealth, respectively.

5. Discussion

5.1. Comparison Based on Machine Learning Methods and Suggestion

Table 3 shows the fitting results of the LightGBM method and the OLS method, which was commonly used in the previous literature for the data set. Based on the goodness of fit and mean absolute error results, we find that LightGBM outperforms the OLS method, which indicates that the effect of choosing LightGBM to improve the traditional linear model is very significant. Therefore, in future related research, we can try to use more machine learning algorithms to process the data to supplement the theoretical scope of the traditional linear model.

5.2. Key Characteristics Affecting Family Wealth and Suggestion

The results on the ranking of the importance of characteristics in Table 4 show that cognition and literacy play an important influence on household wealth, both at the level of the household as a whole and at the level of household members. However, the mechanism of cognition and literacy in influencing household wealth is usually manifested in the rational understanding of financial instruments and rational allocation of financial assets. Therefore, despite the rapid development of digital financial inclusion today, the relevant departments should still make great efforts to popularize financial literacy among the

public, so that the ability of family wealth management and allocation can be enhanced and the overall level of family wealth can be improved.

5.3. Multidimensional Analysis for Digital Inclusive Finance and Suggestion

According to Figures 1–4, we find that the breadth of coverage and depth of use of Digital Inclusive Finance has a significant contribution to household wealth, and the higher the level of development, the more pronounced the promotion effect. Therefore, relevant authorities should continue to boost the coverage area and financial business functions of Digital Inclusive Finance, so that digital technology and inclusive finance can be more deeply integrated. Furthermore, it is important to implement the policy of financial inclusion and promote digital transformation so that more customer groups can participate in financial activities. Finally, we will enhance the public's awareness and acceptance of digital finance and their knowledge, so that more people can enjoy the benefits of digital financial inclusion, and thus increase their income.

Unfortunately, Digital Inclusive Finance has a significant disincentive effect on household wealth in terms of the degree of digitization due to a possible digital divide. Digital Inclusive Finance is overly reliant on internet connectivity and is unable to reach individuals without mobile phones or digital devices. Today, a large number of rural households in China still suffer from barriers to internet access, and many face a greater degree of digital exclusion. The digital divide has a dampening effect on household wealth accumulation, making the contribution of digital inclusion to household wealth weaker. Alleviating the 'digital divide' and achieving the goal of equalization are important elements in the development of Digital Inclusive Finance. In promoting the implementation of digital inclusion, the relevant authorities should consider the balanced characteristics of the hardware, as well as the universal rules of software design. It is also important to ensure data security and avoid information leakage.

6. Conclusions

Digital finance is one of the most influential forms in the development of inclusive finance, and also an important driving force for the development of inclusive finance. In the context of digital penetration, it is important to deeply explore the key factors affecting household wealth accumulation to enhance household property income and achieve wealth preservation and appreciation. Based on this, this paper uses data from the Peking University Digital Financial Inclusion index and China Household Tracking Survey to construct a more comprehensive household wealth prediction index system at both household and family member levels. The LightGBM machine learning method has been used to verify the relationship between Digital Inclusive Finance and household wealth, and a cumulative local effect algorithm is also introduced to multi-dimensionally analyze the impact of a digital inclusive sub-index on the impact of family wealth accumulation. It was found that the machine learning model could avoid the disadvantages of traditional statistical models, because it does not rely on samples to assume a specific functional form, and calculates the degree of influence of variables on target values based on node storage gain values. A machine learning model is more suitable for analyzing the nonlinear and interactive relationships between variables, identifying the contribution of variables, and then mining the data for more complex and reliable patterns. Secondly, the fitting effect of the model adding Digital Inclusive Finance is significantly improved, which indicates Digital Inclusive Finance has a positive impact on the prediction of household wealth; in addition, there is a strong positive relationship between Digital Inclusive Finance and individual knowledge literacy, family size, and member structure. Families with a strong understanding and rich life experience can better promote wealth creation and accumulation. Thirdly, the analysis of the effect of Digital Inclusive Finance on the growth of family wealth finds that the Digital Financial Inclusion index, the breadth of coverage, and the depth of use sub-indices all show a significant positive relationship with family

wealth, while the degree of digitization has a negative impact, indicating that there is a "digital divide" phenomenon.

The shortcomings of this paper are that, on the one hand, the degree of development of digital financial inclusion is changing rapidly. There may be some deviations in the description of the real situation due to the constraints of data availability. On the other hand, this paper only uses one machine learning model method, and the arguments for robustness testing are not sufficient. The comparison of multiple models can be incorporated into the study in the future.

Author Contributions: Conceptualization, Y.L. and H.Z.; methodology, Y.L. and H.Z.; software, Y.L. and H.Z.; validation, Y.L. and H.Z.; formal analysis, Y.L. and J.S.; investigation, Y.L. and H.Z.; resources, Y.L. and H.Z.; data curation, H.Z.; writing—original draft preparation, H.Z.; writing—review and editing, Y.L. and H.Z.; visualization, H.Z.; supervision, J.S. and Y.T.; project administration, J.S. and Y.T.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Social Science Foundation of China, grant number 20BTJ062.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The [family wealth related data e.g., "TotalAssect"] data that support the findings of this study are available from ["China Family Panel Studies (CFPS)"], [hyperlink to data source, "http://www.isss.pku.edu.cn/cfps/download" (accessed on 25 February 2021)]. The [index of Digital Inclusive Finance e.g., "coverage_breadth"] data that support the findings of this study are available from ["Digital Inclusive Finance Index of Peking University"], [hyperlink to data source, "https://idf.pku.edu.cn/zsbz/index.htm" (accessed on 18 February 2021)].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huang, Y.P.; Huang, Z. China's Digital Finance Development: Now and the Future. *China Econ. Q.* **2018**, *17*, 1489–1502. [CrossRef]
2. Yin, Z.C.; Peng, C.Y.; Angela, C.L. The Development and Impact of Financial Inclusion of Household in China. *J. Manag. World* **2019**, *35*, 74–87. [CrossRef]
3. Zhou, T.Y.; Chen, M.X. Digital Penetration, Inclusive Finance and Household Wealth Growth. *J. Financ. Econ.* **2021**, *47*, 33–47. [CrossRef]
4. Wu, Y.; Peng, C.Y.; Yin, Z.C. Financial Literacy, Wealth Accumulation and Household Asset Allocation. *Mod. Econ. Sci.* **2016**, *38*, 19–29.
5. Greenwell Brandon, M.; Boehmke, B.C.; McCarthy, A.J.A. Simple and Effective Model-Based Variable Importance Measure, statistics. *arXiv* **2018**, arXiv:1805.04755.
6. Lei, X.Y.; Zhou, Y.G. On the Portfolio Choice of Chinese Households:Health and Risk Preference. *J. Financ. Res.* **2010**, *1*, 31–45.
7. Xie, M.B. Household Wealth, Education, and Age Equalization of Wealth—A Study of Household Net Worth Determinants Based on CHFS. *Southeast Acad. Res.* **2017**, *5*, 201–219. [CrossRef]
8. Wang, C.; Yao, L.; Chai, S.J. The Influence of Age Structure on Family Asset Allocation and Its Regional Difference. *Stud. Int. Financ.* **2017**, *2*, 76–86. [CrossRef]
9. Lu, X.M.; Zhao, S.; Luo, R.H. Will Regional Financial Development Promote Rational Household Investment?—From the Perspective of Household Asset Portfolios Diversification. *Res. Econ. Manag.* **2019**, *40*, 60–87. [CrossRef]
10. Yu, N.; Wang, Y. Can Digital Inclusive Finance Narrow the Chinese Urban–Rural Income Gap? The Perspective of the Regional Urban–Rural Income Structure. *Sustainability* **2021**, *13*, 6427. [CrossRef]
11. Smith, J.P.; McArdle, J.J.; Willis, R. Financial Decision Making and Cognition in a Family Context. *Econ. J.* **2010**, *120*, F363–F380. [CrossRef] [PubMed]
12. Meng, Y.J. Cognitive ability and household asset selection. *Econ. Res. J.* **2014**, *49*, 132–142.
13. Behrman, J.R.; Mitchell, O.S.; Soo, C.K.; Bravo, D. How Financial Literacy Affects Household Wealth Accumulation. *Am. Econ. Rev.* **2012**, *102*, 300–304. [CrossRef]
14. Hu, Z. Financial Literacy and Household Wealth Accumulation:Based on Micro-data from Urban Household in China. *J. Zhongnan Univ. Econ.* **2018**, *4*, 110–117. [CrossRef]
15. Johnston, D.; Kassenboehmer, S.C.; Shields, M.A. Financial decision-making in the household: Exploring the importance of survey respondent, health, cognitive ability and personality. *J. Econ. Behav. Organ.* **2016**, *132*, 42–61. [CrossRef]

16. Biljanovska, N.; Palligkinis, S. Control thyself: Self-control failure and household wealth. *J. Bank. Financ.* **2018**, *92*, 280–294. [CrossRef]
17. Bernardi, F.; Boertien, D.; Geven, K. Childhood Family Structure and the Accumulation of Wealth Across the Life Course. *J. Marriage Fam.* **2018**, *81*, 230–247. [CrossRef]
18. Chu, C.Y.C.; Kan, K.; Lin, J.C. Variations of wealth resemblance by family relationship types in modern Chinese families. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 6548–6553. [CrossRef]
19. Stein, L.; Yannelis, C. Financial Inclusion, Human Capital, and Wealth Accumulation: Evidence from the Freedman's Savings Bank. *Rev. Financial Stud.* **2020**, *33*, 5333–5377. [CrossRef]
20. Célerier, C.; Matray, A. Bank-Branch Supply, Financial Inclusion, and Wealth Accumulation. *Rev. Financ. Stud.* **2019**, *32*, 4767–4809. [CrossRef]
21. von Fintel, D.; Orthofer, A. Wealth inequality and financial inclusion: Evidence from South African tax and survey records. *Econ. Model.* **2020**, *91*, 568–578. [CrossRef]
22. Duvendack, M.; Mader, P. Impact of financial inclusion in low- and middle-income countries: A systematic review of reviews. *Campbell Syst. Rev.* **2019**, *15*, 594–629. [CrossRef]
23. Chen, L. From Fintech to Finlife: The case of Fintech Development in China. *China Econ. J.* **2016**, *9*, 225–239. [CrossRef]
24. Zhang, Q.; Posso, A. Thinking Inside the Box: A Closer Look at Financial Inclusion and Household Income. *J. Dev. Stud.* **2017**, *55*, 1616–1631. [CrossRef]
25. Li, J.; Wu, Y.; Xiao, J.J. The impact of digital finance on household consumption: Evidence from China. *Econ. Model.* **2019**, *86*, 317–326. [CrossRef]
26. Song, X.L. Empirical Analysis of Digital Inclusive Finance Bridging the Urban-rural Residents' Income Gap. *Financ. Econ.* **2017**, *6*, 14–25.
27. Zhang, D.H.; Yin, Z.C.; Sui, Y.B. Can Financial Inclusion Improve the Quality of Poverty Reduction:Analysis Based on Multidimensional Poverty. *South China J. Econ.* **2020**, *10*, 56–75. [CrossRef]
28. Zhang, B.Q.; Wu, W.T. Digital Inclusive Finance, Entrepreneurship and Income Distribution:An Empirical Research from the Perspective of Urban-rural Differences in China. *Chin. Rev. Financ. Stud.* **2021**, *13*, 31–44.
29. Ji, X.; Wang, K.; Xu, H.; Li, M. Has Digital Financial Inclusion Narrowed the Urban-Rural Income Gap: The Role of Entrepreneurship in China. *Sustainability* **2021**, *13*, 8292. [CrossRef]
30. Li, Y.; Wang, M.; Liao, G.; Wang, J. Spatial Spillover Effect and Threshold Effect of Digital Financial Inclusion on Farmers' Income Growth—Based on Provincial Data of China. *Sustainability* **2022**, *14*, 1838. [CrossRef]
31. Zhao, H.; Zheng, X.; Yang, L. Does Digital Inclusive Finance Narrow the Urban-Rural Income Gap through Primary Distribution and Redistribution? *Sustainability* **2022**, *14*, 2120. [CrossRef]
32. Song, J.; Zhou, H.; Gao, Y.; Guan, Y. Digital Inclusive Finance, Human Capital and Inclusive Green Development—Evidence from China. *Sustainability* **2022**, *14*, 9922. [CrossRef]
33. Xie, X.L.; Zhang, H.X.; Guo, F. Can Digital Finance Promote Entrepreneurship?—Evidence from China. *China Econ. Q.* **2018**, *17*, 1157–1580. [CrossRef]
34. Zhang, X.; Wan, G.H.; Zhang, J.J.; He, Z.Y. Digital Economy, Financial Inclusion, and Inclusive Growth. *Econ. Res. J.* **2019**, *54*, 71–86. [CrossRef]
35. Jia, L.; Tan, W.; Abu, M. Financial Literacy, Family Wealth and Family Entrepreneurial Decision-Making. *Southwest Financ.* **2021**, *1*, 83–96.
36. Sun, Y.; Zhang, T.; Wang, X.; Li, D. Status, Problems and Prospects of Digital Inclusive Finance in China. *Quant. Technol. Econ.* **2021**, *38*, 43–59. [CrossRef]
37. Fu, Q.Z.; Huang, Y.P. The Heterogeneous Impact of Digital Finance on Rural Financial Needs—Evidence from China Household Finance Survey and Peking University Digital Inclusive Finance Index. *J. Financ. Res.* **2018**, *11*, 68–84.
38. Hu, L.; Yao, S.Q.; Yang, C.Y.; Ji, L.H. Is digital financial inclusion good for alleviating relative poverty? *J. Financ. Econ.* **2021**, *47*, 93–107. [CrossRef]
39. Yu, C.; Jia, N.; Li, W.; Wu, R. Digital inclusive finance and rural consumption structure—Evidence from Peking University digital inclusive financial index and China household finance survey. *China Agric. Econ. Rev.* **2021**, *14*, 165–183. [CrossRef]
40. Ye, Y.; Pu, Y.; Xiong, A. The impact of digital finance on household participation in risky financial markets: Evidence-based study from China. *PLoS ONE* **2022**, *17*, e0265606. [CrossRef]
41. Hong, Y.; Wang, S. Big Data, Machine Learning and Statistics: Challenges and Opportunities. *China J. Econom.* **2021**, *1*, 17–35.
42. Araújo, A. Polynomial regression with reduced over-fitting—The PALS technique. *Measurement* **2018**, *124*, 515–521. [CrossRef]
43. Adner, R.; Puranam, P.; Zhu, F. What Is Different About Digital Strategy? From Quantitative to Qualitative Change. *Strat. Sci.* **2019**, *4*, 253–261. [CrossRef]
44. Nambisan, S.; Lyytinen, K.; Majchrzak, A.; Song, M. Digital Innovation Management: Reinventing Innovation Management Research in a Digital World. *MIS Q.* **2017**, *41*, 223–238. [CrossRef]
45. Athey, S.; Imbens, G.W. The State of Applied Econometrics: Causality and Policy Evaluation. *J. Econ. Perspect.* **2017**, *31*, 3–32. [CrossRef]
46. Chernozhukov, V.; Chetverikov, D.; Demirer, M.; Duflo, E.; Hansen, C.; Newey, W.; Robins, J. Double/debiased machine learning for treatment and structural parameters. *Econ. J.* **2018**, *21*, C1–C68. [CrossRef]

47. Sezer, O.B.; Ozbayoglu, A.M. Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach. *Appl. Soft Comput.* **2018**, *70*, 525–538. [CrossRef]
48. Hu, W.X.; Su, Z.X.; Yang, L. Research on the Prediction Effect of Investor Attention to the Concept Index on the Re-turn of Concept Index and the Investment Strategy of Concept Index Based on the Random Forest Algorithm. *Front. Sci. Technol. Eng. Manag.* **2021**, *40*, 60–66.
49. Malhotra, S.; Reus, T.H.; Zhu, P.; Roelofsen, E.M. The Acquisitive Nature of Extraverted CEOs. *Adm. Sci. Q.* **2017**, *63*, 370–408. [CrossRef]
50. Lu, Y.; Zhang, Y.Q.; Li, H.; Zhao, H.Y. Managerial individual characteristics and corporate performance: Evidence from a machine learning approach. *J. Manag. Sci. China* **2020**, *23*, 120–140.
51. Wen, B.H.; Gou, S.D.; Zhang, B.R. Evaluation of Liquidity Monitoring Methods of China's Commercial Banks in the Process of Leverage Adjustment: Comparison of Algorithms Based on Deep Feedforward Networks. *Front. Sci. Technol. Eng. Manag.* **2019**, *38*, 59–66.
52. Chen, D.M.; Wang, L.Z.; Chen, A.N. Digitalization and Strategic Management Theory:Review, Challenges and Pro-spects. *J. Manag. World* **2020**, *36*, 220–236. [CrossRef]
53. Li, J.; Zhu, X.Q.; Wang, L.X. Partners Selection in the Cooperative Innovation of Cluster Enterprises Based on Knowledge-associated Evaluation Index System and Improved ELMAN Neural Network Model. *Front. Sci. Technol. Eng. Manag.* **2020**, *39*, 23–29.
54. Hurst, E.; Lusardi, A. Liquidity Constraints, Household Wealth, and Entrepreneurship. *J. Politi Econ.* **2004**, *112*, 319–347. [CrossRef]
55. Gai, Q.E.; Zhu, X.; Shi, X.H. The Heterogeneous Effect of Wealth on Entrepreneurship: An Empirical Analysis Based on Farmers in Three Provinces. *J. Financ. Econ.* **2013**, *39*, 134–144. [CrossRef]
56. Gan, Y.; Zhu, J.; Liu, C.Y. The study on Impacts of Financial Constraints towards Household Entrepreneurship and its Differences between Urban and Rural Areas. *Shanghai J. Econ.* **2015**, *9*, 15–23. [CrossRef]
57. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3146–3154. [CrossRef]
58. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2018**, *267*, 1–38. [CrossRef]
59. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 22071–22080. [CrossRef]
60. Apley, D.W.; Zhu, J. Visualizing the effects of predictor variables in black box supervised learning models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2020**, *82*, 1059–1086. [CrossRef]

sustainability

MDPI

Article

A Full Population Auditing Method Based on Machine Learning

Yasheng Chen, Zhuojun Wu * and Hui Yan

Department of Accounting, School of Management, Xiamen University, Xiamen 361005, China
* Correspondence: anthony604329469@163.com

Abstract: As it is urgent to change the traditional audit sampling method that is based on manpower to meet the growing audit demand in the era of big data. This study uses empirical methods to propose a full population auditing method based on machine learning. This method can extend the application scope of the audit to all samples through the self-learning feature of machine learning, which helps to address the dependence on auditors' personal experience and the audit risks arising from audit sampling. First, this paper demonstrates the feasibility of this method, then selects the financial data of a large enterprise for full population testing, and finally summarizes the critical steps of practical applications. The study results indicate that machine learning for full population auditing is able to detect, in all samples, abnormal business whose execution does not adhere to existing accounting rules, as well as abnormal business with irregular accounting rules, thus improving the efficiency of internal control audits. By combining the learning ability of machine-learning algorithms and the arithmetic power of computers, the proposed full population auditing method provides a feasible approach for the intellectual development of future auditing at the application level.

Keywords: artificial intelligence; fraud detection; audit analytics; machine learning

1. Introduction

Essentially, an audit is a process of continuously collecting, identifying, and comprehensively applying audit evidence. Limited by cost-effectiveness and technical methods, current audit procedures are always based on audit sampling. This results in the neglect of many specific business samples and audit opinions based on the sample results, thus hiding huge audit risks [1]. With the continuous innovation of technology, it has become possible to change the characteristics of audit work from "partial" and "accurate" to "complete" and "efficient" [2–6]. Performing full population auditing allows auditors to obtain a much-increased level of audit evidence, which provides a higher level of assurance for the audit conclusion [1].

With the electronization of financial and business data associated with the audited entity, auditors can obtain more comprehensive data, and audit tasks can be gradually structured and repeated. Auditors can analyze the data with the assistance of technical tools to form more scientific audit conclusions [5,7]. However, faced with a large amount of accumulated audit data from the auditees, the research progress of full-population auditing is still laborious. One reason is that the audit data are of a huge volume and have high complexity. Based on the current audit method, the computer only performs tasks according to the input instructions or rules but cannot understand and learn any audit knowledge during the audit process. The completion of the audit work is still dependent on the personal experience of the auditors, which makes the audit efficiency limited by manpower. Another reason is that current big data analysis is still limited to specific auditees or application scenarios. Though many current technical tools can, to a certain extent, improve the efficiency of data storage, data retrieval, and data research, it is necessary to establish different analysis rules for varying analysis requirements that cannot be extended and

applied to all business scenarios. Therefore, it is imperative and challenging to implement full population testing.

Given the difficulties of full population auditing, this paper proposes to conduct full population auditing by using machine learning. Machine learning is a new artificial intelligence technology that can learn autonomously and summarize the hidden rules in data [8]. Different from the way that people set rules for machines, the self-learning feature of machine learning can significantly reduce the dependence on personal experience and help overcome the workforce limitation in traditional audit work [9]. Furthermore, machine learning is not limited by specific scenarios. In the face of different types of business data, machine learning can learn and summarize hidden business rules [10]. Therefore, this paper proposes a full population auditing method based on machine learning. This method includes two functions. The first function is to help the machine to master the accounting rules used by the financial staff by constructing a machine-learning model. The second function is to enable the machine to identify anomalies in all accounting records. Specifically, based on the constructed machine-learning model, the machine should be able to simulate the role of a human in accounting processing. Then, by comparing the accounting results derived from the machine and those that were manually derived, abnormal accounting records can be identified.

2. Research Background and Literature Review

In the modern risk-oriented audit framework, sampling risk refers to the possibility that the auditor's conclusions, based on the sampling results, do not conform to those obtained from the same audit procedure performed by the audit object. Non-sampling risk refers to the possibility that the auditor will make incorrect conclusions due to factors other than the sample size [11]. Though statistical sampling has many inherent drawbacks, it is the only method that should be adopted under limited conditions, and any sampling method will produce sampling risks. If the auditors conduct full population testing, the risks of audit sampling can be effectively avoided [1].

In a previous study, researchers attempted to improve audit efficiency in theory and practice. From a theoretical point of view, sampling audit is the universal choice for modern audit methods. As a cutting-edge tool for data analysis, machine learning emphasizes the combination of computing power and the learning ability of various algorithms to analyze massive data. Many researchers believe that machine learning is significant for abnormal audit discovery and judgment [5,7,12–14]. Huang et al. have introduced a method of applying audit data analysis and machine learning to full population testing and have discussed the feasibility of this method [1].

In practice, many researchers have attempted to obtain more effective audit information from mass data by introducing technical tools, based on the improvement of computing power and the development of big data audit. Issa et al. [3] have pointed out that new technologies such as cloud computing and Hadoop can store and process data more efficiently. Deep learning technology can automatically extract features from unstructured or semi-structured data, such as images, voice, text, video, etc. Appelbaum et al. [15] have proposed that many analysis techniques such as regression techniques, descriptive statistics, expert systems, big data visualization [16], and other big data analysis techniques have been widely used in audit tasks to cope with the increasing requirements for audit data analysis. Huang and Vasarhelyi [17] have proposed an RPA framework for audit scenarios to help auditors perform audit tasks with high repeatability and low judgment requirements. To sum up, researchers have paid much attention to the collation and analysis of massive audit data. Still, the audit process is dependent on the personal experience of the auditors, and the problem of the limits placed on audit work by human resources has not been fundamentally solved. Machine learning has made progress in many areas, such as optimizing management needs [18], mental health monitoring [19] and other areas but there are also some applications of machine learning in auditing. Lokanan et al. have proposed the use of the Mahalanobis distance to calculate the deviation distance of

financial reports, identify abnormal financial reports, and divide them into different credit ratings [20]. Kotsiantis et al. have proposed the use of the financial ratio to build a decision tree model to classify financial statements in order to identify false financial statements [21].

However, in the business audit of the accounting process of an enterprise, the question of how to apply machine learning to full population auditing has been less studied and there is also a lack of practical application cases.

In view of the above literature, this paper collates Table 1 and refines the point of view of each reference.

Table 1. Overview of related literature.

Author	Research Conclusion	Limitation
Huang et al. [1]	With the development of audit data analysis and machine learning, full population auditing is feasible. Through the priority examination of abnormal or suspicious transactions, the effectiveness of the audit can be improved.	Lack of implementation of a full population auditing framework
No W G et al. [7]	This study proposes a multi-dimensional audit data selection (MADS) framework, which provides a systematic method for auditors to deal with a large number of potential outliers in the process of audit data selection.	The risk brought by sampling audit is not solved in essence.
Issa et al. [3]	Scanning and OCR, electronic recording and computing, cloud technology, blockchain, intelligent contract, big data storage can all be used as the technical basis, and their technical application in audits can be divided into sensing, archiving and prediction technology, meta-control, exogenous measurement, rapid detection of phenomena, integration of evidence, and deep learning data in the audit.	Lack of discussion on the necessity of application in audit
Appelbaum et al. [15]	This paper introduces the research progress of big-data-related technology in audit, which breaks the view that the audit major only focuses on ratio analysis, sampling and scanning.	Lack of research progress in application
Feiqi Huang; Miklos A. Vasarhelyi [17]	This paper introduces RPA into audit practice and puts forward a four-stage framework for guidance.	This platform requires modification when extended to other platforms, and the services of third-party websites need to be used in practical applications.
Lokanan et al. [20]	Proposes the use of the Mahalanobis distance to calculate the deviation distance of financial reports, identify abnormal financial reports, and divide them into different credit ratings.	It is difficult to determine the threshold level of outliers.
Kotsiantis et al. [21]	Proposes the use of the financial ratio to build a decision tree model to classify financial statements in order to identify false financial statements.	Only two-classification scenarios are discussed, and no multi-classification audit scenarios are involved.

3. Research Question

In financial auditing, auditors can exhaust the transactions over the significance threshold and sample the transactions under the importance threshold to obtain audit opinions. In order to enhance the rationality of the audit opinion, this study attempts to cancel the audit sampling. For the proposed, machine-learning-based, full population auditing method, this study first demonstrates the feasibility of the method, and then selects the data of a large state-owned enterprise from 2020 to 2021 to further test and analyze the feasibility of the method. Finally, this study summarizes the key steps of practical application.

Full population auditing requires a review of each accounting record to identify unusual misstatements. To implement full population testing, it is necessary to know the accounting rules of various businesses and strictly check whether each business is properly accounted for according to the rules. If full population auditing is conducted using personal experience, a lot of human resources are consumed, and human errors cannot be avoided.

With the improvement in computing power, computers can use machine-learning algorithms to autonomously learn massive data, summarize hidden laws with far greater efficiency than human beings and identify anomalies in the data. "Learning" refers to a process in which, for a class of tasks T and performance metrics P, a computer program improves its performance on tasks T as measured by performance metrics P through experience E [22]. Machine learning first determines the specific algorithm according to the task;. It then divides the preprocessed data set into a training set and a test set and trains the model by using the training set. Meanwhile, it selects the optimal function expression of the model by calculating the loss function. Finally, it measures the generalization performance of the optimal function on the test set. This self-learning characteristic of machine learning enables computers to have, to a certain extent, human-like professional judgment, thus removing the limitation imposed on the analysis process by its dependence on personal experience. Moreover, the application of machine learning is not limited to a specific type of sample data. Instead, it can be applied to various scenes to find out the rules according to the input data and complete the corresponding classification or prediction tasks. Machine learning relies on different algorithms to solve data problems, and the specific algorithm depends on the type of problem to solve. Common algorithms include supervised learning, unsupervised learning, reinforcement learning, etc. An appropriate machine-learning algorithm can be chosen according to the needs of different scenarios [23].

The stronger the regularity between data, the better the learning effect of machine learning. Since financial datasets have uniform accounting rules and standards, machine learning has natural application advantages in processing financial data [5]. In the audit scenario, machine learning can learn and summarize the hidden rules from the huge and complex audit data accumulated by the audited unit and identify any anomalies in the audit data by establishing a machine-learning model [12,13]. Compared with the traditional audit mode, machine learning adopts a completely different approach: it changes from relying on personal experience to training computer programs to master accounting rules, and it makes audit judgments by exploiting sufficient computing power. Thus, it can quickly and accurately locate abnormal accounting treatments. Different from the risk-oriented audit model, which is mainly aimed at specific businesses [24], machine learning makes it possible to perform full population auditing while avoiding audit risks caused by human factors. Moreover, machine learning is limited to a specific audited unit or a specific audit scenario. Owing to its self-learning characteristic, it can process various input audit data and so can improve the performance of various applications. The above analysis indicates that the characteristics of machine learning can meet the needs of full population testing, which indicates that this method is feasible from a theoretical point of view.

On the basis of the above feasibility analysis, this study aims to propose a full population auditing method based on machine learning rather than on manpower and so overcome the dependence of an audit on manpower and thus the risks associated with sampling audits. In order for the machine to audit, two targets need to be achieved. The first target is to enable the machine to master the accounting rules used by financial personnel, and the second target is to enable the machine to detect anomalies in accounting processing based on the mastered accounting rules.

In order to achieve the research target, this study selected the travel expenses of an enterprise and tried to implement full population auditing based on machine learning at the application level. According to the above discussion, the process of full population auditing can be divided into two stages. First of all, it is necessary to build a machine-learning model based on existing accounting records. By learning accounting rules hidden in the accounting records, the machine can simulate the accounting process of financial personnel.

Then, this study can use the machine-learning model to account for all travel expenses, and after comparing the results provided by the machine with the manual accounting records, we can quickly identify abnormal samples. In this way, this study implements this machine-learning-based full population auditing method.

The main research question of this study is whether one can use a machine to effectively replace the manual work needed to achieve effective full population testing. Referring to the two functions proposed above, it can be divided into the following two research questions:

Research Question 1: Can the machine master the accounting rules in the enterprise's financial data based on machine learning?

Research Question 2: Can the constructed machine-learning model accurately screen out abnormal accounting treatments?

4. Research Design

Based on the above feasibility study, this study selected a large state-owned enterprise in China as the implementation object, one whose business data are of many types, of large quantities, and with high repeatability, and formally tested the effectiveness of this full population auditing method in practice. Specifically, a machine-learning-based full population audit was conducted on travel expenses data from 2020 to 2021. In the bookkeeping process, financial staff need to analyze the business information of travel expenses and record them to the correct accounting subject. The full population auditing method can test whether financial personnel record all travel expenses to the correct account in the bookkeeping process and can support more sufficient audit conclusions based on the more comprehensive audit evidence obtained.

The full population auditing process was divided into two stages. In the first stage, a machine-learning model was constructed based on the financial data by selecting sample data to train the model and determining the accounting rules hidden in the data. In the second stage, the constructed model was used to check the existing accounting treatment and detect the anomalies in the accounting records [6].

Specifically, in the first stage, in order to enable the machine to learn the existing accounting rules in accounting processing, this study constructed a machine-learning model based on the travel expense sample in 2020. At this stage, 80% of the samples of travel expenses in 2020 were randomly divided into training sets, and the remaining 20% into test sets. The machine learned the accounting rules according to the randomly selected training set samples and used the test set to test the validity of the model. According to the case scenario, this study selects the appropriate machine-learning algorithm training model and selects appropriate evaluation indicators to evaluate the performance of the model, so as to analyze whether the machine-learning model has mastered the existing accounting rules. In addition, in the case of a sufficient number of samples, if the machine cannot accurately learn the accounting rules of a certain type of business, it shows that there are anomalies in this kind of business sample, which may be due to unclear design of the accounting rules or other abnormal reasons. In the second stage, in order to test whether the machine can identify abnormal samples in accounting processing, this study took all travel expenses in 2021 as test samples to test the practicability of the machine-learning model. In this study, the machine-learning model constructed in the first stage was used to account for each sample of travel expenses in 2021. The machine can determine which type of business the sample should be classified into. Each type of business corresponds to a different account. By comparing the account given by the machine and the account given by the financial staff, we can know whether the two results are consistent. From the perspective of a financial audit, this study took the consistency between machine accounting processing results and manual accounting processing records as the evaluation standard. If there was an inconsistency between the two results, then the sample was recognized as abnormal. For example, we can use the machine-learning model on a regular basis or after each bookkeeping, output the machine's judgment of the business sample, and then match with the existing manual accounting records. If the match succeeds, the

samples are considered to be processed correctly. If the match fails, the business sample is considered to be at risk and is detected as a fraud sample.

In this way, the machine-learning-based full population auditing method can quickly locate abnormal samples. Specifically, it enables the machine to master the accounting rules of financial personnel, which can effectively improve the audit efficiency. By using machines instead of manpower, machines can examine more samples with far more efficiency than manpower. This method can not only help to detect the anomalies in the existing accounting rules, but also help enterprises to check the human errors in the business. For large-scale enterprises in particular, this method can help enterprises to locate small errors quickly and obtain more sufficient audit evidence with a lower audit cost.

4.1. Data Sources and Sample Selection

This study selected 14,681 travel expense samples in 2020 to construct a machine-learning model and applied the model to conduct full population auditing of 10,738 travel expense samples in 2021. The business data and accounting records of the travel expenses in this study were all obtained from the internal database of the enterprise. As shown in Table 2, each travel expense sample included transaction information from nine business features, based on which business staff needs to determine which of the 16 debit accounts should be accounted for. Additionally, each business feature is a classification variable and contains several categories. For example, the feature variable "Document Type" contains five different categories of transaction information in total. Each debit account represents a category of travel expenses (the accounting records of travel expenses include debit account, credit account and amount. In this case, the credit account is automatically generated by the system according to the payment method, which is usually an account such as a "bank deposit" or "cash", and the amount is automatically filled by the system, which has nothing to do with the specific business, so this study chooses the debit account to represent the business category).

Table 2. The statistics of the travel expense business sample data set in 2020.

Feature Variable	Categories Quantity	Debit Account
Document Type	5	Administrative Expenses-A
Level-1 Budget Usage Department	12	Administrative Expenses-B
Level-2 Budget Usage Department	44	Administrative Expenses-C
Budget Project	8	Main Business Cost-A
Budget Activity	11	Main Business Cost-B
Budget Item	24	Main Business Cost-C
Currency Type	13	Main Business Cost-D
Travel Type	11	Main Business Cost-E
Training Type	13	Main Business Cost-F Payroll Payable-A Payroll Payable-B Payroll Payable-C Selling Expenses-A Selling Expenses-B Other Business Costs-A Long-term Deferred Expenses-A

Note 1: When Level-1 accounts are the same, different types of debit accounts are distinguished by the suffixes A, B, and C.

In order to enable machines to replace humans to undertake bookkeeping, it is necessary to clarify the input and output of the model. In this case, "input" refers to the information in nine business features of travel expenses, and "output" refers to the accounting subject that needs to be recorded. By summarizing the rules in the data, the machine-learning model can imitate human beings to establish mapping rules between the "input" and "output".

The travel expense sample dataset in 2020 involves nine feature variables and sixteen corresponding debit accounts. The details of the sample dataset are presented in Table 2.

Table 3 presents the sample distribution of each debit account in the travel expense dataset in 2020.

Table 3. The statistics of the travel expense business distribution in 2020.

Debit Account	Sample Size	Percentage	Debit Account	Sample Size	Percentage
Administrative Expenses-A	2991	20.37%	Main Business Cost-F	559	3.81%
Administrative Expenses-B	50	0.34%	Payroll Payable-A	10	0.07%
Administrative Expenses-C	1	0.01%	Payroll Payable-B	38	0.26%
Main Business Cost-A	8119	55.3%	Payroll Payable-C	196	1.34%
Main Business Cost-B	40	0.27%	Sales Expenses-A	2233	15.21%
Main Business Cost-C	53	0.36%	Sales Expenses-B	42	0.29%
Main Business Cost-D	343	2.34%	Other Business Costs-A	1	0.01%
Main Business Cost-E	4	0.03%	Long-term Deferred Expenses-A	1	0.01%

In Table 3, there are 16 types of travel expense businesses in 2020. Among these, the three types of travel expenses represented by "Administrative Expenses-A", "Main Business Cost-A", and "Sales Expenses-A" involve 13,343 samples in total, accounting for 90.89% of the total sample. Meanwhile, the number of each type of travel expense business corresponding to "Administrative Expenses-C", "Main Operation Cost-E", "Payroll Payable-A", "Other Business Costs-A", and "Long-term Deferred Expenses-A" is ten or less. Furthermore, the five types of business involve 17 samples, accounting for no more than 0.12% of the total sample, and they belong to the debit account rarely found in travel expenses. To cover all types of travel expense businesses, this study retained the entire travel expense business sample in 2020.

To thoroughly test the effectiveness of the model, this study excluded travel expense business added in 2021, such as the travel expense business involving a new debit account or the travel expense business involving a new department and an existing debit account. Then, 10,738 samples of travel expense business (excluding new business) in 2021 were used to conduct full population auditing, as shown in Table 4.

Table 4. The statistics of the travel expense business sample data set in 2021.

Feature Variable	Categories Quantity	Debit Account
Document Type	4	Administrative Expenses-A
Level 1 Budget Usage Department	8	Main Business Cost-A
Level 2 Budget Usage Department	40	Main Business Cost-B
Budget Project	8	Main Business Cost-C
Budget Activity	11	Main Business Cost-D
Budget Item	16	Main Business Cost-F
Currency Type	3	Payroll Payable-A
Travel Type	10	Payroll Payable-B
		Payroll Payable-C
		Selling Expenses-A
Training Type	13	Selling Expenses-B
		Other Business Costs-A
		Long-term Deferred Expenses-A

Note 2: In the travel expenses business in 2021, there was no business related to the three debit items: "Administrative Expenses-B", "Administrative Expenses-C", and "Main Business Cost-E".

The sample distribution of each debit account in the 2021 travel expense dataset is presented in Table 5 (as the percentage of each type of sample size is calculated with only two digits retained, the sum of percentages is not strictly equal to 100%).

Table 5. The statistics of the travel expense business distribution in 2021.

Debit Account	Sample Size	Percentage	Debit Account	Sample Size	Percentage
Administrative Expenses-A	2300	21.42%	Payroll Payable-B	28	0.26%
Main Business Cost-A	6630	61.74%	Payroll Payable-C	107	1.00%
Main Business Cost-B	165	1.54%	Sales Expenses-A	795	7.4%
Main Business Cost-C	6	0.06%	Sales Expenses-B	7	0.07%
Main Business Cost-D	209	1.95%	Other Business Costs-A	1	0.01%
Main Business Cost-F	475	4.42%	Long-term Deferred Expenses-A	9	0.08%
Payroll Payable-A	6	0.06%			

In Table 5, the travel expense business in 2021 involves 13 debit accounts, and each account represents a type of travel expense business. Among these, the three types of businesses "Administrative Expenses-A", "Main business cost-A", and "Sales Expenses-A" involve 9733 samples, accounting for 90.47% of the total sample. The number of samples of "Main Business Cost-C", "Payroll Payable-A", "Sales Expenses-B", "Other Business Costs-A", and "Long-term Deferred Expenses-A" are all less than 10. The five types of businesses involve a total of 31 samples, accounting for no more than 0.28% of the total sample.

4.2. Machine-Learning Model Selection

To establish a machine-learning model suitable for the audit scenario of this study, the machine-learning algorithm is chosen from the following four aspects: First, this study can obtain the input business data and the output accounting results, so a supervised learning algorithm is more suitable. Secondly, as the model's output is a discrete class variable, this is considered a classification problem. Thirdly, when the prediction results of the model are inconsistent with the accounting records, the judgment process of the model needs to be verified, so the selected algorithm should be interpretable [25].

The decision tree algorithm is one of the few supervised learning algorithms with good interpretability in machine learning [26]. There are three classical decision tree algorithms: the ID3, C4.5, and CART. Among these, the CART decision tree algorithm that is based on the two former algorithms is proposed, which makes up for their defects [27]. Based on the above considerations, this study adopts the CART decision tree algorithm to establish a machine-learning model.

A CART classification regression tree model was proposed by Breiman et al. in 1984 [28]. CART consists of feature selection, tree generation and pruning, and can be used for both classification and regression. A CART algorithm uses a binary recursive segmentation technique, which divides the current sample into two subsample sets, so that the generated non-leaf nodes have two branches. Therefore, CART is actually a binary tree. When CART is a classification tree, a *Gini* value is used as the basis for splitting nodes, and when CART is used as a regression tree, the minimum variance of samples is used as the basis for splitting nodes. For the binary classification problem, if the probability that the sample points belong to the first category is p, then the *Gini* index of the probability distribution is:

$$Gini(p) = 2p(1-p)$$

For a given sample set D, the *Gini* index is:

$$Gini(D) = 1 - \sum_{k=1}^{K} \frac{|C_k|^2}{|D|}$$

where C_k is a subset of samples belonging to class k in D and K is the number of classes. $|C_k|$ and D indicate the number of subsets and the number of samples, respectively.

If the sample set D is divided into D_1 and D_2 according to whether feature A takes a possible value a, that is:

$$D_1 = \{(x, y) \in D | A(x) = a\}, D_2 = D - D_1$$

Under the condition of feature A, the Gini index of set D is:

$$Gini(D, A) = \frac{D_1}{D} Gini(D_1) + \frac{D_2}{D} Gini(D_2)$$

The *Gini* index *Gini*(*D*,*A*) denotes the uncertainty of the set after the decomposition of $A = a$. The greater the *Gini* index, the greater the uncertainty of the sample set.

The steps to build the CART model are:

(1) For each feature A dataset D can be divided into dataset D_1 and D_2 according to whether the sample point A is a or not. Calculate the *Gini* index of the existing feature for the data set.
(2) For all features A and all possible segmentation points a, the features with the lowest *Gini* index and the corresponding segmentation points are selected as the optimal features and segmentation points.
(3) Step (1) (2) is called recursively on the optimal subtree until the stop condition is met.
(4) Generate the CART classification tree.

According to the tree structure, it divides the complex decision-making process into a series of simple choices and can intuitively show the whole decision-making process. The decision tree starts from a root node, passes through a series of branch nodes, and finally obtains a plurality of leaf nodes. This process reflects the mapping relationship between the input data and output results. Considering the travel expense sample as the root node, each branch representative from the top to bottom subdivides the sample according to a specific category. Then, according to evaluation indicators such as the Gini coefficient, each level of subdivision category can be filtered until the travel expense sample is subdivided to the last leaf node level by level with the optimal path. The construction process of the decision tree is similar to the behavior pattern of people when they are making decisions. Its foundation is the establishment of the correspondence between the input data and output results through a series of rules. According to the above introduction, the CART decision tree algorithm is suitable for learning accounting rules in enterprises. This paper attempts to build a full population auditing model based on a CART classification tree (the model parameters are set as default values (the evaluation index of the model used in this paper is the Gini coefficient, the selection criterion of the split point is set as global optimum, and there is no limit to the number of categories to be considered. If a branch node is subdivided, the number of samples for that node should not be less than one, and the Gini coefficient must be greater than 0. The maximum number of nodes of the same branch is set to 6, and the minimum number of samples of leaf nodes is 2, which does not limit the maximum number of samples of leaf nodes. When dividing the sample data set, the model divides the training set and the test set in a random way (criterion="gini",splitter="best",max_depth=None,min_samples_split=2,min_samples_leaf=1, min_weight_fraction_leaf=0.0,max_features=None,random_state=None,max_leaf_nodes= None,min_impurity_decrease=0.0,ccp_alpha=0.0)), and use this model to test the accounting process of all samples.

4.3. Process Implementation

Step 1: Filter relevant business feature variables.

As mentioned above, based on the analysis of the accounting rules formulated by the enterprise, the following nine business feature variables were selected as inputs to the model: (1) Document Type; (2) Level-1 Budget Usage Department; (3) Level-2 Budget Usage Department; (4) Budget Project; (5) Budget Activity; (6) Budget Item; (7) Currency Type; (8) Travel Type; (9) Training Type.

Step 2: Data preprocessing.

For the dataset in this study, missing values need to be handled, and feature variables should be converted into feature vectors.

For the missing values in the dataset, this study considered them from the aspect of business features. Specifically, the missing values in each business feature variable are separated into one category and regarded as business feature variables. For example, the "Training Type", a business feature variable, has null values in some samples. Additionally, null values are considered as another category. In addition to the other 12 categories with content in the "Training Type", the "Training Type" includes 13 categories in total.

As for converting feature variables into feature vectors, since the corresponding categories in the business features are all text information, to enable the machine to understand the relationship between different categories of feature variables, one-hot encoding was adopted to convert the text information into data format that could then be identified by the machine [29]. Taking the feature variable "Budget Item" as an example, this business feature includes eight categories, all of which are text information. After one-hot encoding, these eight categories are converted into one-hot vectors, such as [1, 0, 0, 0, 0, 0, 0, 0], [0, 1, 0, 0, 0, 0, 0, 0], [0, 0, 1, 0, 0, 0, 0, 0], [0, 0, 0, 1, 0, 0, 0, 0], [0, 0, 0, 0, 1, 0, 0, 0], [0, 0, 0, 0, 0, 1, 0, 0], [0, 0, 0, 0, 0, 0, 1, 0], or [0, 0, 0, 0, 0, 0, 0, 1]. To ensure that each category of the feature variables corresponds to a unique one-hot vector, the datasets of the travel business in 2020 and 2021 need to be encoded in the same way (for the sample data set used to construct the full-sample audit model and the sample data set used to implement the full-sample audit, it is necessary to perform one-hot encoding based on the same coding rules to ensure consistency among different sample data sets).

Step 3: Establish a machine-learning model.

Of the travel expenses samples in 2020, 80% were randomly selected as the training set and 20% as the test set, and the CART decision tree model was established. By analyzing the model indicators, the effectiveness of the model can be evaluated.

Step 4: Conduct full population auditing of the new business samples.

After constructing the CART decision tree model, the data of travel expenses in 2021 were input into the model, and the debit accounts were predicted in turn. In this process, inconsistent accounting treatments can be screened out by comparing the debit account predicted by the model with the existing accounting records.

5. Results and Analysis

To answer research question 1: "Can the machine master the accounting rules in the enterprise's financial data based on machine learning?", in the third step of building the machine-learning model, this study chose the CART decision tree algorithm and used the travel expense business dataset in 2020 to train and construct the CART decision tree model. The evaluation indicator results of the model on the test set are listed in Table 6.

In this study, the model's performance can be evaluated by the following evaluation indicators: precision, recall, F1 score, and accuracy. Specifically, precision = TP/(TP + FP), where TP indicates the number of correctly predicted samples in a business type, and the sum of TP and FP is the number of samples predicted as this business type in the test set. This measures the percentage of the samples correctly predicted as a business type. Recall = TP/(TP + FN), where the definition of TP is the same as above, and the sum of TP and FN indicates the number of the business type's samples in the test set. This measures the percentage of samples correctly predicted in a business type. The F1 score is a comprehensive indicator that considers both precision and recall. Furthermore, accuracy measures the overall classification accuracy.

The results indicate that the precision and recall of nine types of travel expenses in the test set are not less than 99%, and the F1 scores all reach 1. This part of the business samples accounts for 98.37% of the samples in the test set, indicating that the model achieves excellent prediction results for these nine types of business samples. However, for the two types of businesses corresponding to "Employee Remuneration Payable-B" and

"Employee Remuneration Payable-C", the precision of the model is 85% and 80%, the recall is 85% and 94%, and the F1 score is 0.85 and 0.86, respectively. This indicates that only 85% and 84% of the samples predicted to be these two types of businesses were correctly classified, and only 85% and 94% of the samples of these two types of businesses were correctly classified, respectively. The classification results of the model have significantly decreased in these two types of business samples. Additionally, only four samples and ten samples were included for the businesses of "Main Business Cost-E" and "Employee Remuneration Payable-A" respectively, which resulted in a sparse distribution of samples. As a result, the model cannot provide enough information to learn the correct accounting rules, so the model cannot classify these two types of businesses well.

Table 6. The evaluation indicator results of the CART decision tree model.

Debit Account	Precision	Recall	f1 Score	Support
Administrative Expenses-A	1.00	1.00	1.00	621
Administrative Expenses-B	1.00	1.00	1.00	11
Main Business Cost-A	1.00	0.99	1.00	1596
Main Business Cost-B	1.00	1.00	1.00	8
Main Business Cost-C	1.00	1.00	1.00	13
Main Business Cost-D	1.00	1.00	1.00	58
Main Business Cost-E	0.00	0.00	0.00	1
Main Business Cost-F	1.00	1.00	1.00	112
Payroll Payable-A	0.00	0.00	0.00	1
Payroll Payable-B	0.85	0.85	0.85	13
Payroll Payable-C	0.80	0.94	0.86	50
Sales Expenses-A	1.00	1.00	1.00	444
Sales Expenses-B	1.00	1.00	1.00	9
Accuracy			0.99	2937
macro-average	0.82	0.83	0.82	2937
micro-average	0.99	0.99	0.99	2937

Note 3: Precision = TP/(TP + FP), where TP is the number of correctly predicted samples in a business type, and the sum of TP and FP is the number of samples predicted as this business type in the test set. Recall = TP/(TP + FN), where the definition of TP is the same as above and the sum of TP and FN is the number of samples of this business type in the test set. F1Score = 2 * Precision * Recall/(Precision + Recall). Accuracy = (TP + TN)/(TP + TN + FN + FP), where the sum of TP and TN is the number of correctly predicted samples in the test set, and the sum of TP, TN, FN, and FP is the number of samples in the test set. Macro-average refers to the arithmetic average of the indicator. Micro-average refers to the weighted average of the indicators. The weight is the percentage of the number of test samples in each type over the total number of test samples. Support refers to the number of business samples. Note 4: The number of samples in each of the three debit accounts "Administrative Expenses-C", "Other Business Costs-A", and "Long-term Deferred Expenses-A" is one. As the samples are too sparse to divide the training set and test set normally, the classification results only cover the remaining 13 accounts.

In view of the fact that the decision tree is an explainable model, we can understand its classification process of travel expenses from the perspective of machines. As the first research aim includes enabling the machine to master the accounting rules used by financial personnel, this study makes the machine simulate the way of thinking of financial personnel by constructing the decision tree model. By explaining the specific decision-making process of the decision tree model, we can know whether the machine masters the accounting rules. In the decision tree model, each category of business characteristic can become the branch point of the decision tree. Here is an example to describe the classification process of the decision tree model. For example, according to the decision tree model that was constructed, in this case, for a travel sample, and without knowing the business result, the machine performs the following judgment process: (1) check whether "Budget Activity" is "training expense"; (2) check whether "Budget Activity" is "marketing expense"; (3) check whether "Budget Project" is "promotion project" (4) check whether "Level 1 Budget Usage Department" is "Beijing branch"; (5) check whether "Budget Item" is "publicity speech"; (6) check whether "Level 2 Budget Usage Department" is "business unit"; (7) check whether "Travel Type" is "business trip"; (8) check whether "Training Type" is "advanced training"; (9) check whether "Document Type" is "domestic travel reimbursement"; (10) check

whether "Budget Item" is "daily business"; and (11) check whether "Currency Type" is "RMB". After all the above checks, it is concluded that the sample should be classified into this category of business with debit account "Administrative Expenses-A". As in this example, for research question 1, "can the machine master the accounting rules in the enterprise's financial data based on machine learning", if the machine's accounting results of the sample are consistent with the manual accounting records, it means that the machine has mastered the accounting rules of the financial staff in the sample.

The analysis results indicate that the model has mastered the accounting rules for the nine business samples mentioned above, which account for the vast majority (due to the existence of a small number of human errors in the travel business in 2020, the precision for some business types rate or recall rates were 99%, which did not reach one. In the case of large samples, this does not affect the overall prediction results of the model). For the two types of businesses of "Payroll Payable-B" and "Payroll Payable-C", where both the accuracy and recall had significantly decreased, it was found through the manual review that these two types of businesses had not formed clear accounting rules in their accounting treatment, and that there is subsequently confusion and errors in judgment. Specifically, these two types of businesses are travel expense businesses for staff training. However, the relevant accounting rules are not apparent when conducting accounting treatment, and the classification of these types of businesses is often judged based on the personal experience of financial personnel. For example, as shown in Table 7, Sample 1 and Sample 2 are travel expense businesses for staff training in 2020. The two samples have the same feature variables, but Sample 1 is debited to the account "Payroll Payable-C", and Sample 2 is debited to the account "Payroll Payable-B".

Table 7. Example 1 of travel expenses sample.

Business Feature	Business Record	
	Sample 1	**Sample 2**
Year	2020	2020
Document Type	1	1
Level 1 Budget Usage Department	7	7
Level 2 Budget Usage Department	39	39
Budget Project	3	3
Budget Activity	2	2
Budget Item	1	1
Currency Type	1	1
Travel Type	3	3
Training Type	3	3
Debit Account (Accounting record)	Payroll Payable-C	Payroll Payable-B

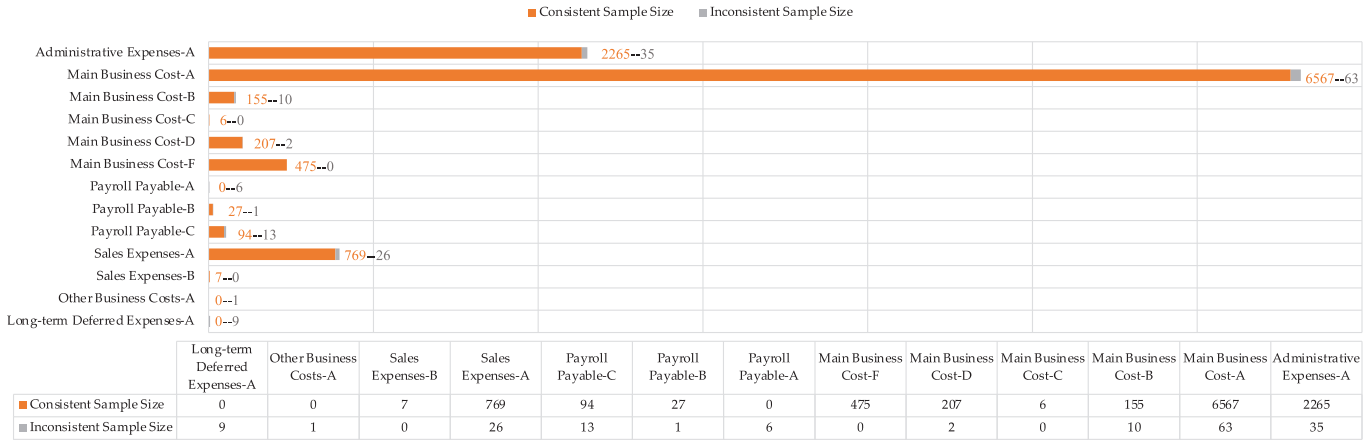
Note 5: Due to the inconvenience to enterprises associated with the disclosure of their data, the categories of business feature variables are represented by Arabic numerals, and numbers do not represent the size.

As for whether the accounting rules in the enterprise financial data can be obtained by employing the machine-learning method raised in research question 1, the above analysis indicates that the machine can obtain the accounting rules from the business data with clear rules by using the machine learning method. However, the machine cannot learn clear accounting rules from the business data with unclear business rules. The unclear formulation of accounting rules often reveals defects in the design of an enterprise's internal controls, while machine learning can also effectively identify such anomalies.

To answer research question 2: "Can the constructed machine-learning model accurately screen out the existing abnormal accounting treatments?", in the fourth step, full population auditing of the travel expense samples in 2021 was conducted by using the constructed CART decision tree model. Then, the prediction results of the model were compared with the existing accounting records, and the statistical results are presented in Table 8 and Figure 1.

Table 8. The comparison between the accounts predicted by the model and the accounts in the accounting record of travel expense business in 2021.

Debit Account	Sample Size	Consistent Sample Size	Percentage of Consistent Samples	Inconsistent Sample Size	Percentage of Inconsistent Samples
Administrative Expenses-A	2300	2265	98.48%	35	1.52%
Main Business Cost-A	6630	6567	99.05%	63	0.95%
Main Business Cost-B	165	155	93.94%	10	6.06%
Main Business Cost-C	6	6	100.00%	0	0.00%
Main Business Cost-D	209	207	99.04%	2	0.96%
Main Business Cost-F	475	475	100.00%	0	0.00%
Payroll Payable-A	6	0	0.00%	6	100.00%
Payroll Payable-B	28	27	96.43%	1	3.57%
Payroll Payable-C	107	94	87.85%	13	12.15%
Sales Expenses-A	795	769	96.73%	26	3.27%
Sales Expenses-B	7	7	100.00%	0	0.00%
Other Business Costs-A	1	0	0.00%	1	100.00%
Long-term Deferred Expenses-A	9	0	0.00%	9	100.00%
Grand Total	10,738	10,572	98.45%	166	1.55%



	Long-term Deferred Expenses-A	Other Business Costs-A	Sales Expenses-B	Sales Expenses-A	Payroll Payable-C	Payroll Payable-B	Payroll Payable-A	Main Business Cost-F	Main Business Cost-D	Main Business Cost-C	Main Business Cost-B	Main Business Cost-A	Administrative Expenses-A
Consistent Sample Size	0	0	7	769	94	27	0	475	207	6	155	6567	2265
Inconsistent Sample Size	9	1	0	26	13	1	6	0	2	0	10	63	35

Figure 1. The comparison between the accounts predicted by the model and the accounts in the accounting record of travel expenses in 2021.

The results show that the consistency between the prediction results of the constructed model and the existing records reaches 98.45%, indicating that most business samples have been accounted for following the original rules. In particular, among the 10,738 travel business samples in 2021, the model prediction results of 166 business samples are inconsistent with the actual accounting records. These samples were accurately located after model screening and required a critical manual review.

This study reviewed the above 166 samples according to business types. The review results are shown in Table 9. Taking into account the inconsistency between the machine-learned and the manual results, and bearing in mind that there may be unclear accounting rules, there are three types of situations. In the first type of situation, the machine prediction result is correct, but the manual bookkeeping result is incorrect. In the second situation, the bookkeeping record is correct, but the machine prediction result is incorrect. In the third situation, the accounting rules are unclear, and it is not easy to judge whether the manual and the machine are correct. Next, these three types of samples are analyzed with examples.

In the first situation, there are 116 samples, accounting for 69.88% of all inconsistent samples, which indicates that most errors in real business are caused by manual failure to follow the original accounting rules. For example, as shown in Table 10, Sample 1 and Sample 2 have the same feature variables. The former was debited to the account "Main Business Cost-A" in 2020, and the latter was debited to the account "Sales Expenses-A"

in 2021, indicating that Sample 2 was not accounted for following the original accounting rules.

Table 9. The manual review results of inconsistent samples.

Debit Account	Inconsistent Sample Size	Manual Review Results		
		1. Machine Predictions Are Correct	2. Unclear Accounting Rules	3. Accounting Records Are Correct
Administrative Expenses-A	35	35	0	0
Main Business Cost-A	63	16	47	0
Main Business Cost-B	10	10	0	0
Main Business Cost-D	2	2	0	0
Payroll Payable-A	6	6	0	0
Payroll Payable-B	1	0	0	1
Payroll Payable-C	13	12	0	1
Sales Expenses-A	26	26	0	0
Other Business Costs-A	1	0	0	1
Long-term Deferred Expenses-A	9	9	0	0
Grand Total	166	116	47	3

Table 10. Example 2 of travel expenses sample.

Business Feature	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Year	2020	2021	2020	2020	2020	2020
Document Type	1	1	1	1	1	1
Level 1 Budget Usage Department	10	10	7	7	7	7
Level 2 Budget Usage Department	7	7	27	27	12	37
Budget Project	1	1	3	3	3	3
Budget Activity	1	1	2	2	2	2
Budget Item	1	1	1	1	1	1
Currency Type	1	1	1	1	1	1
Travel Type	1	1	1	1	3	3
Training Type	1	1	1	1	4	3
Debit Account (Accounting record)	Main Business Cost-A	Sales Expenses-A	Main Business Cost-A	Payroll Payable-C	Main Business Cost-A	Administrative Expenses-A
Debit Account (Machine Prediction)	–	Main Business Cost-A	–	–	–	–

In the second situation, there are 47 samples, accounting for 28.31% of all inconsistent samples. These samples were misjudged because the accounting rules were not formulated clearly, and the model could not learn clear rules. The 47 samples in 2021 involving the debit account of "Main Business Cost-A" belong to the same business (If the values of the feature variables are the same, the business is considered the same). Through the manual review, it was found that the business had unclear accounting rules in 2020. In 2020, there were 24 samples of this business, as shown in sample 3 in Table 10. Among these samples, eight were accounted as the debit account of "Main Business Cost-A", while the other 16 samples were accounted as the debit account of "Payroll Payable-C", as shown in Sample 4. Therefore, the model could not learn accurate accounting rules, and there were relevant abnormal samples in 2021. Additionally, it should be noted that although there were 6,630 business samples involving the debit account of "Main Business Cost-A" in 2020, accounting for 61.74% of the total number of samples, and the samples with unclear accounting rules only accounted for 0.7% of this business type, it was difficult to find the anomaly through the evaluation indicators with the established model. However, by predicting each business sample in 2021, the problem can be accurately located.

In the third situation, there are three samples, accounting for 1.81% of all abnormal samples, which is an accidental situation. After a manual review, as for the sample corresponding to the debit account of "Other Business Costs-A", the model cannot correctly predict this type of business because there is only one similar sample in 2020, and the sample

fails to provide sufficient information. For the remaining two samples corresponding to the debit account of "Payroll Payable-B" and "Payroll Payable-C", the reason for the incorrect prediction is that the machine learned from the human error in the same business in 2020, so the machine predicted the samples according to incorrect rules. These two abnormal samples, i.e., Samples 5 and 6, were erroneously accounted for as the debit account of "Main Business Cost-A" and "Administrative Expenses-A" in 2020, and they were correctly accounted for as the debit account of "Employee Remuneration Payable-B" and "Employee Remuneration Payable-C" in 2021, thus causing the inconsistency in this sample.

As for the question of whether the machine-learning model established in research question 2 can accurately screen out abnormal accounting treatment, the result analysis indicates that by implementing the full population auditing method, the abnormal samples in the enterprise business can be accurately located. In this study, this method can detect misstated samples that are not accounted for following the original rules and further accurately detect business samples that have unclear accounting rules. Therefore, the proposed method provides a supplement for the detection of internal control problems for enterprises in the model building stage and avoids situations in which these problems remain undetected due to the relatively small proportion of abnormal samples. Additionally, if there had been occasional human errors in the enterprise business in previous years, the proposed audit method based on machine learning can accurately locate those anomalies.

6. Conclusions, Limitations and Further Research

This paper proposes a full population auditing method that combines the learning ability of a machine-learning algorithm and a computer's computing power. The effectiveness of full population auditing based on machine learning has been proven in practice, providing an essential reference for the intellectual development of future auditing. By taking the data of a state-owned enterprise as an example, this paper selected 14,681 travel expense samples in 2020 to construct a CART decision tree model and applied the model to conduct a full population audit of 10,738 travel expense samples in 2021. The research results indicate that full population auditing based on machine learning can accurately detect and locate the following two types of anomalies: misstatements in the accounting treatment of enterprise business, and the abnormal business in which the accounting rules are not formulated clearly. It is worth noting that the proposed method can also identify the anomalies for some low-frequency business samples. Therefore, the proposed full population auditing method based on machine learning can help enterprises quickly review all business data and accurately locate misstatements in the business. Additionally, this method can detect internal control problems of businesses in a timely manner and enable enterprises to improve the quality of their financial data and the effectiveness of internal control management.

For theoretical implications, this study demonstrates the feasibility of the full population auditing method with machine learning. Meanwhile, the relevant literature has been enriched by the proposal for a framework by which a full population auditing method can be used.

For practical implications, this study expands the application scenarios of machine learning in the audit field, as it applies a machine-learning-based full population auditing method to the accounting record audit scene of actual enterprises. Additionally, this study verifies the application value of the method through an enterprise case by effectively identifying abnormal situations in the design and implementation of internal control. Furthermore, this method can accurately locate the position of misstatements in big data and detect very small misstatements. This method can break through the limitations that result from the way in which risk-oriented audits pay more attention to significant misstatements, and it promotes the further development of full population auditing research. Moreover, this study provides a necessary reference for other audit application scenarios.

This study is limited to the discussion of full population auditing in the bookkeeping process. Due to the high threshold of professional knowledge, there may not be enough professional personnel to understand and develop its practical applications. For future

research, the machine learning method used in this paper can also be extended to identify other internal control audit issues, including internal control design and internal control execution. The design defects of internal control will inevitably lead to inconsistent handling behaviors of different personnel. Additionally, the lack of internal control implementation effectiveness will hinder the effective execution by authorized personnel of the corresponding processes according to the due rules. From the perspective of application scenarios, future research will also explore the use of the full population auditing method in other business scenarios.

Based on the above, a general process of the proposed full population auditing method based on machine learning is presented below.

- Step 1: Filter relevant business feature variables.

To select the business feature variables to construct the model, it is necessary to fully understand the business processing logic of the financial personnel to ensure that the selected feature variables provide sufficient rule information and help construct an effective machine-learning model. For example, in the above study, nine business features, such as "Document Type" and "Level-1 Budget User Department" were selected for the travel business to provide sufficient accounting rule information for the subsequent model construction.

- Step 2: Data preprocessing.

When constructing data sets, data preprocessing is usually needed to facilitate the subsequent construction of the machine-learning model. Specifically, it includes a series of contents, such as processing missing values and data transcoding. For missing values, this study regards the missing value of each business feature as a category value or deletes samples with too many missing values. For data transcoding, the sample data are converted into a data format that the machine can recognize by feature discretization (feature discretization is a process of dividing feature values into multiple categories by using segmentation points) or one-hot encoding. For example, in the above research, the text information in the sample data is converted into one-hot vectors for machine recognition.

- Step 3: Construct a machine-learning model.

The construction of a machine-learning model involves selecting an appropriate machine-learning algorithm according to the application scene, dividing the sample dataset used for constructing a model into a training set and a test set at a particular proportion, constructing a machine-learning model, and using the dataset to train the model.

- Step 4: Conduct full population auditing of the new business samples.

After preprocessing the sample dataset, the constructed machine-learning model is used to predict the sample to be audited. By comparing the prediction results of the model with the existing records, the inconsistent business samples are further reviewed.

Author Contributions: Conceptualization, Z.W.; Methodology, Z.W.; Validation, Y.C. and H.Y.; Writing—review & editing, Y.C. and H.Y.; Supervision, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, grant number 72172132.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huang, F.; No, W.G.; Vasarhelyi, M.A.; Yan, Z. Audit data analytics, machine learning, and full population testing. *J. Financ. Data Sci.* **2022**, *8*, 138–144. [CrossRef]
2. Deloitte. Deloitte Forms Alliance with Kira Systems to Drive the Adoption of Artificial Intelligence in the Workplace. 2016. Available online: https://www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/deloitte-forms-alliance-with-kira-systems-to-drive-theadoption-of-artificial-intelligence-in-the-workplace.html (accessed on 30 October 2022).
3. Issa, H.; Sun, T.; Vasarhelyi, M.A. Research ideas for artificial intelligence in auditing: The formalization of audit and workforce supplementation. *J. Emerg. Technol. Account.* **2016**, *13*, 1–20. [CrossRef]
4. PricewaterhouseCoopers (PwC). Blockchain and Smart Contract Automation: How Smart Contracts Automate Digital Business. 2016. Available online: https://www.pwc.com/us/en/technology-forecast/2016/blockchain/pwc-smart-contract-automation-digital-business.pdf (accessed on 30 October 2022).
5. Kokina, J.; Davenport, T.H. The emergence of artificial intelligence: How automation is changing auditing. *J. Emerg. Technol. Account.* **2017**, *14*, 115–122. [CrossRef]
6. Chan, D.Y.; Vasarhelyi, M.A. Innovation and practice of continuous Auditing. In *Continuous Auditing (Rutgers Studies in Accounting Analytics)*; Chan, D.Y., Chiu, V., Vasarhelyi, M.A., Eds.; Emerald Publishing Limited: Bingley, UK, 2018; pp. 271–283.
7. No, W.G.; Lee, K.; Huang, F.; Li, Q. Multidimensional audit data selection (MADS): A framework for using data analytics in the audit data selection process. *Account. Horiz.* **2019**, *33*, 127–140. [CrossRef]
8. Zhou, Z.H. *Machine Learning*; Springer Nature: Berlin/Heidelberg, Germany, 2021.
9. Reavie, V. Do You Know the Difference between Data Analytics and AI Machine Learning? 2018. Available online: https://www.forbes.com/sites/forbesagencycouncil/2018/08/01/do-you-know-the-difference-between-data-analytics-and-ai-machine-learning/?sh=3347ef185878 (accessed on 30 October 2022).
10. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef] [PubMed]
11. *Auditing Standards 2315*; Audit Sampling; Public Company Accounting Oversight Board (PCAOB): Washington, DC, USA, 2014.
12. Rapoport, M.; Auditing Firms Count on Technology for Backup. Wall Street Journal. 2016. Available online: https://www.wsj.com/articles/auditing-firms-count-on-technology-for-backup-1457398380?mod=djem_jiewr_AC_domainid (accessed on 30 October 2022).
13. Brennan, B.; Flynn, M.; Baccala, M. Artificial Intelligence Comes to Financial Statement Audits. CFO.com. 2 February 2017. Available online: http://www.cfo.com/auditing/2017/02/artificial-intelligence-audits/ (accessed on 30 October 2022).
14. Dickey, G.; Blanke, S.; Seaton, L. Machine Learning in Auditing. The CPA Journal. 2019. Available online: https://www.cpajournal.com/2019/06/19/machine-learning-in-auditing/ (accessed on 30 October 2022).
15. Appelbaum, D.A.; Kogan, A.; Vasarhelyi, M.A. Analytical procedures in external auditing: A comprehensive literature survey and framework for external audit analytics. *J. Account. Lit.* **2018**, *40*, 83–101. [CrossRef]
16. Keim, D.; Qu, H.; Ma, K.L. Big-data visualization. *IEEE Comput. Graph. Appl.* **2013**, *33*, 20–21. [CrossRef] [PubMed]
17. Huang, F.; Vasarhelyi, M.A. Applying robotic process automation (RPA) in auditing: A framework. *Int. J. Account. Inf. Syst.* **2019**, *35*, 100433. [CrossRef]
18. Alhasnawi, B.N.; Jasim, B.H.; Sedhom, B.E.; Guerrero, J.M. Consensus Algorithm-based Coalition Game Theory for Demand Management Scheme in Smart Microgrid. *Sustain. Cities Soc.* **2021**, *74*, 103248. [CrossRef]
19. Dadi, K.; Varoquaux, G.; Houenou, J.; Bzdok, D.; Thirion, B.; Engemann, D. Population modeling with machine learning can enhance measures of mental health. *GigaScience* **2021**, *10*, giab071. [CrossRef] [PubMed]
20. Lokanan, M.; Tran, V.; Vuong, N.H. Detecting anomalies in financial statements using machine learning algorithm: The case of Vietnamese listed firms. *Asian J. Account. Res.* **2019**, *4*, 181–201. [CrossRef]
21. Kotsiantis, S.; Koumanakos, E.; Tzelepis, D.; Tampakas, V. Predicting fraudulent financial statements with machine learning techniques. In Proceedings of the Hellenic Conference on Artificial Intelligence, Heraklion, Greece, 18–20 May 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 538–542.
22. El Naqa, I.; Murphy, M.J. What is machine learning? In *Machine Learning in Radiation Oncology*; Springer: Cham, Switzerland, 2015; pp. 3–11.
23. Mahesh, B. Machine learning algorithms—A review. *Int. J. Sci. Res.* **2020**, *9*, 381–386.
24. Liu, J. The enterprise risk management and the risk oriented internal audit. *Int. Bus. J.* **2012**, *10*, 287–292. [CrossRef]
25. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine learning interpretability: A survey on methods and metrics. *Electronics* **2019**, *8*, 832. [CrossRef]
26. Du, M.; Liu, N.; Hu, X. Techniques for interpretable machine learning. *Commun. ACM* **2019**, *63*, 68–77. [CrossRef]
27. Mienye, I.D.; Sun, Y.; Wang, Z. Prediction performance of improved decision tree-based algorithms: A review. *Procedia Manuf.* **2019**, *35*, 698–703. [CrossRef]
28. Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. Cart. Classification and Regression Trees. 1984. Available online: https://imsarchives.nus.edu.sg/oldwww/Programs/014swclass/files/mark.pdf (accessed on 30 October 2022).
29. Yang, K.K.; Wu, Z.; Bedbrook, C.N.; Arnold, F.H. Learned protein embeddings for machine learning. *Bioinformatics* **2018**, *34*, 2642–2648. [CrossRef] [PubMed]

sustainability

MDPI

Article

Financial Fraud Detection of Listed Companies in China: A Machine Learning Approach

Yasheng Chen and Zhuojun Wu *

Department of Accounting, School of Management, Xiamen University, Xiamen 361005, China
* Correspondence: anthony604329469@163.com

Abstract: As the focus of capital market supervision, financial report fraud has shown a development trend of enormous numbers, complex transactions, and hidden means in recent years. To improve audit efficiency and reduce the dependence on non-financial data, the study only uses the structured original data in the financial report to constructs a new fraud identification model, which can quickly detect fraud in China. This study takes the listed companies in China from 1998 to 2016 as research samples and selects 28 sets of raw data from financial reports. Then, this study compares the detection effectiveness of two single classification machine learning algorithms and five ensemble learning algorithms on fraud detection. Compared with single classification machine learning algorithms, the results show that ensemble learning algorithms are generally better at detecting fraud for Chinese listed companies, and the stacking algorithm performs the best. The study results provide direct evidence for rapid fraud detection using financial report raw data and ensemble learning algorithms. The study first proposes a stacking algorithm-based financial reporting fraud identification model for listed companies in China, which provides a simple and effective approach for investors, regulators, and management. It can also provide a reference for the detection of other fraud scenarios.

Keywords: artificial intelligence; machine learning; fraud detection; financial report; financial disclosures; China; securities fraud

1. Introduction

Financial report fraud has occurred frequently in recent years. Due to the increasingly hidden fraud of listed companies, traditional fraud detection methods cannot meet the review requirements promptly. Limited by human resources and detection methods, it usually takes several months, or even years, for listed companies to obtain an administrative penalty notice from the China Securities Regulatory Commission (CSRC) for suspected financial report fraud. "The ACFE Global Fraud Investigation Report" is the most cited source of fraud and job-related crime data in the world. It has been published every two years since it was first published in 1996. The questionnaire was distributed to fraud investigators and anti-fraud practitioners in various countries worldwide, and ACFE sets up a special team to collect, analyze, summarize, and explain the survey. The survey results released in 2020 are based on 2504 survey reports from 125 countries. According to the 2020 ACFE Global Fraud Investigation Report [1], the average investigation time for each fraud case is 14 months. Financial report fraud has led to the highest economic loss for all types of fraud, with a median loss of USD 964,000. Regulatory authorities can effectively reduce the fraud of listed companies through timely intervention and construct a timely and effective fraud identification model for the financial report of listed companies, which is significant for investors, analysts, and regulators [2].

To improve audit efficiency, researchers introduce machine learning as an effective fraud detection tool. Machine learning can find hidden rules in massive data far more efficiently than manual work [3]. Additionally, the detection of financial fraud can be regarded as a classification task in machine learning [4,5]. Generally, machine learning

algorithms can be divided into traditional single classifiers and ensemble classifiers [6–8]. Specifically, the single classifier refers to the classification or prediction of samples by a single classification method. The ensemble classifier integrates the prediction results of multiple single classifiers based on the single classification to obtain the final prediction result. Different ensemble learning algorithms have been developed according to different integration processes [6–8].

However, in previous studies on listed companies in China, researchers only paid attention to how to select appropriate financial ratios or non-financial information to construct a fraud identification model, and hold the view that these financial ratios can cover the complete financial information, but they ignored the direct information in financial reports [9–11]. During the generation process, essential original information may be lost by financial ratios in the financial report [10], and non-financial information originates from different channels. Non-financial information may come from management news, board meeting minutes, contract details, etc. [12]. This information can be divided into three categories, including management characteristics and control environments, governance and organization structures, and relation and industry conditions [5]. For example, management characteristics and control environments include board mobility, highly complex transactions, excessive incentives to management, ineffective accounting, and information systems, etc. Governance and organization structures include the proportion of the largest shareholder, board of directors, board of supervisors, the proportion of independent directors, etc. [13]. Relation and industry conditions include significant related-party transactions not in the ordinary course of business, significant declines in increasing business failures in the industry, etc. The information is basically in an unstructured form, which is difficult to deal with in a unified and centralized way, thus increasing the difficulty of financial report audits [14]. This paper suggests that fraud detection from the financial reporting raw accounting numbers may be more practical and provide a simple and direct approach.

To fully utilize the vital information in raw accounting data, reduce the dependence on non-financial information, and achieve rapid fraud detection in financial reporting, it is necessary to develop a simple and effective fraud detection model using the raw financial data of listed companies in China. This paper selects the financial report raw data of listed companies in China from 1998 to 2016 and tests the applicability of the financial report fraud detection model developed by Bao et al. in China's capital market. Then, based on the research background of listed companies in China, two single classifiers and four ensemble classifiers, namely logistics, SVM, random forest, RUSBoost, XGBoost, AdaBoost, and stacking, are selected to construct financial reporting fraud identification models respectively; meanwhile, comparative research is conducted to determine financial reporting fraud identification models suitable for the capital market in China. The results indicate that the financial report fraud recognition model developed by Bao et al. does not achieve the expected effect in China's capital market. After further comparing the detection effects of various machine learning methods, it is found that the fraud detection effect of the ensemble learning method is generally better than that of the single classifier method. Especially, the financial reporting fraud identification model based on the stacking algorithm achieves the best detection effect, and its AUC reaches 0.742.

The contributions of this paper are summarized as follows. First, the study extracts raw data from financial reports to provide a new perspective to quickly detect fraud in Chinese listed companies from financial reports, improving audit efficiency and reducing the dependence on non-financial data. Then, this paper tests the applicability of the existing western financial reporting fraud identification model in the Chinese capital market. Subsequently, the study uniformly compares the effectiveness of two single classifiers and five ensemble classifiers in detecting fraud of listed companies in China; the results provide direct evidence for rapid fraud detection using financial report raw data and ensemble learning algorithms. Finally, under the background of listed companies in China, the study first proposes a stacking algorithm-based financial reporting fraud identification model

for listed companies in China, which provides a simple and effective detection method for investors, analysts, and regulators.

2. Literature Review

There is some important literature related to variable selection and method selection. Researchers have tried to construct an effective financial reporting fraud detection model to identify financial reporting fraud more efficiently and accurately, by using financial indicators and machine learning methods.

2.1. Raw Accounting Data and Financial Ratios

Researchers at home and abroad have paid much attention to the selection of financial ratios to establish fraud identification models. For example, Skousen (2008) [15] selected five pressure indicators and two opportunity indicators according to the fraud triangle theory and established a logistic regression model to evaluate the impact of each financial indicator on financial fraud. Additionally, Spathis (2002) [16], Kirkos (2006) [17], Ravisankar (2011) [18], and Kanapickienė (2015) [19], used financial ratios and non-financial data for fraud detection. Aiming at the raw financial report data, Bao et al. (2020) [10] used 28 sets of raw financial report data to construct a RUSBoost model that can effectively detect the fraud of listed companies in the United States, in which 28 groups of detailed raw data can be referred to the details of Section 4. However, few studies have proposed a fraud detection model for listed companies in China based on the original financial report data. Therefore, this paper puts forward Hypothesis 1: the RUSBoost model applicable to the listed companies in the US can achieve the same forecast effect in China.

2.2. Single Classifier and Ensemble Classifier

In the existing research, the single classifier algorithm such as logistics and support vector machine (SVM) is widely used to detect financial fraud of listed companies in China [4,9,19,20], and the adopted ensemble algorithm includes random forest, XGBoost, etc. [14,21,22]. Nagai (2010) [23] surveyed 49 related studies published from 1997 to 2008 and found that the single classifier method was still the mainstream in financial fraud detection research during this period. Logistic regression is a classic single-classifier method in fraud detection research [9]. Many researchers have also proposed to use ensemble algorithms for fraud detection. Kotsiantis (2006) [24] proposed to use several machine learning methods to integrate the stacking model to construct a fraud prediction model to predict financial fraud. Hajeck's (2017) [25] research uses the AdaBoost algorithm to classify financial fraud. Baesens (2021) [26] believes that the XGBoost algorithm is suitable for fraud detection. Bao et al. (2020) [10] applied the RUSBoost algorithm to construct a fraud detection model suitable for the US capital market.

However, there has been little research on using raw financial reporting data to construct machine learning models for listed companies in China, and there is no uniform comparison of the effectiveness of different machine learning methods. The algorithms taken for comparison in this study include two traditional machine learning methods (the logistics model and SVM) and four ensemble learning algorithms (random forest, AdaBoost, XGBoost, RUSBoost, and stacking). To construct a financial reporting fraud detection model suitable for the Chinese market, this paper compares the detection effects of single classifiers and ensemble classifiers and puts forward Hypothesis 2: based on the unified sample data set, the ensemble classifier can achieve better fraud detection results than the single classifier.

3. Overview of Machine Learning

Logical regression is often used in classification problems, and its essence is linear regression. The maximum likelihood estimation method is often employed to solve each parameter, and a layer of Sigmoid function is added to the feature-to-result mapping [27,28]. SVM is a supervised binary classification algorithm. Its general idea is to find a hyperplane with

the best generalization ability in the sample space to divide the samples, thus maximizing the distance between the sample points closest to the hyperplane in the two types of samples [29,30]. Compared with a single classifier using a single method, the ensemble classifier makes classifications by combining the results of multiple classifiers. According to the different integration approaches, the ensemble classifier can be divided into three categories: bagging, boosting and stacking [31]. Bagging sets the result of each classifier to the same weight, and the final classification result is obtained by voting, which is represented by the random forest algorithm [32,33]. Bagging gives different weights to different classifiers, and it allocates more weight to mis-divided samples in the classifier's training process, thus improving the performance of the classifier. This type of ensemble classifier includes AdaBoost [34,35], XGBoost [36,37], RUSBoost [38,39], and other variants. Stacking is an ensemble learning technology that integrates multiple classifications models through a meta-classifier. The classifiers selected by stacking are of different types. It first obtains the results of multiple classifiers, then takes the results as features, and uses a meta-classifier, such as logical regression, for training to obtain the final classification results [40,41].

In view of the complexity of the mathematical formulas of these algorithm models, Table 1 is attached to list some relevant literature on each algorithm in detail.

Table 1. Related literature list of algorithm model.

Model	Related Literature
Logistic Regression	Peng C Y J, Lee K L, Ingersoll G M. [27]. Bewick V, Cheek L, Ball J. [28]
SVM	Jakkula V. [29] Chen Y W, Lin C J. [30]
RUSBoost	Seiffert C, Khoshgoftaar T M, Van Hulse J, et al. [38] Seiffert C, Khoshgoftaar T M, Van Hulse J, et al. [39]
AdaBoost	Wu X, Kumar V, Ross Quinlan J, et al. [34] Schapire R E. [35]
Random Forest	Breiman L. [32] Biau G, Scornet E. [33]
XGBoost	Brownlee J. [36] Chen T, Guestrin C. [37]
Stacking	Ting K M, Witten I H. [40] Sill J, Takács G, Mackey L, et al. [41]

4. Verification of Applicability of the RUSBoost Model

4.1. Sample Selection and Data Sources

This paper uses the research sample of all listed companies from 1990 to 2019. The sample data are obtained from the CSMAR financial reporting database and violation database. Given the completeness and accessibility of the data, the financial data involved in this study is derived from the annual report. The CSMAR violations database covers 16 types of violations, as shown in Table 2. Most of the financial reporting frauds of listed companies in China focus on the whitewashing and manipulation of profits [42], so this study selects corporate annual reports with fictitious profits as fraud samples and manually collates 400 corporate annual reports involving fictitious profits, according to the announcement of irregularities by CSMAR.

Table 2. The type of violation for the CSMAR violation database.

Violation Category	
(1) Fictitious profit	(9) False disclosure (other)
(2) Fictitious assets	(10) insider dealing
(3) False records (misleading statements)	(11) Fraudulent listing
(4) Improper handling of general accounting	(12) Illegal trading of shares
(5) Deferred disclosure	(13) Changing the use of funds without authorization
(6) Major omissions	(14) Illegal guarantee
(7) Violation of capital contribution	(15) Manipulate stock prices
(8) Occupation of corporate assets	(16) Other

According to the CSMAR's violation database, this study statistically collates the proportion of the number of companies involved in fictitious profits in the annual financial report from 1990 to 2019 to the total number of listed companies in the database for the current year, as shown in Table 3. The analysis result indicates that according to the regulations of the Ministry of Finance of China, enterprises are required to stop preparing the statement of changes in financial position from 1998 and replace it with the statement of cash flows. Meanwhile, the announcement of financial report fraud is lagging. The proportion of fictitious profit companies announced from 2017 to 2019 to the total number of listed companies in the current year decreased by 0.69%, 0.35%, and 0.10%, respectively, indicating that there may be some companies whose fictitious profit behaviors have not been discovered during this period. Therefore, to ensure the sample's reliability, this study selects companies' annual financial reports from 1998 to 2016 as the sample data, with a total of 35,574 samples, including 337 samples of annual fraud of fictitious profit companies. The data show that between 1998 and 2016, the number of companies involved in fictional profits accounted for about 1% of all listed companies, and the specific proportion fluctuated around 1% each year.

Note that this paper labels listed companies as fraud according to the CSRC's fictional profit penalty announcement, and it is assumed that CSRC has successfully identified all false profit companies, and there is no misstatement in the announcement that has been issued. Correspondingly, in the fraud model built by Bao, there is the same assumption about the samples of American-listed companies. In practice, it is impossible to determine the correct extent to which CSRC can identify fraud, but this does not weaken the importance of the data or the conclusions of this study. For example, assuming that the model can quickly and effectively detect 1% of listed companies with fraud labels, even if fraudulent listed companies account for 2% of all listed companies, it can still save substantial time and help to reduce the risk of investors.

4.2. Selection of Raw Financial Data

Referring to the raw financial data selection by Bao et al. (2020), this paper constructs 28 characteristic variables, of which 25 variables are derived from the raw financial report data; the remaining three variables are "annual turnover of tradable shares", "closing price at the end of the year", and "market value of tradable shares at the end of the year", and they are derived from the CSMAR stock trading database. Due to the differences between some financial reporting items of China and the United States, the corresponding variables in China's capital market and their calculation formulas are listed in Table 4.

Table 3. The statistical distribution of annual reports of fictitious profit companies from 1990 to 2019.

Year	Number of Listed Companies with Fictitious Profits	Total Number of Listed Companies	Percentage of Listed Companies with Fictitious Profits
1990	0	11	0.00%
1991	0	17	0.00%
1992	0	78	0.00%
1993	0	229	0.00%
1994	0	361	0.00%
1995	0	389	0.00%
1996	7	605	1.16%
1997	17	819	2.08%
1998	14	929	1.51%
1999	12	1030	1.17%
2000	17	1175	1.45%
2001	22	1258	1.75%
2002	23	1319	1.74%
2003	24	1381	1.74%
2004	21	1469	1.43%
2005	11	1464	0.75%
2006	12	1547	0.78%
2007	7	1661	0.42%
2008	9	1715	0.52%
2009	7	1864	0.38%
2010	12	2218	0.54%
2011	19	2452	0.77%
2012	22	2580	0.85%
2013	26	2624	0.99%
2014	27	2739	0.99%
2015	41	2927	1.40%
2016	36	3221	1.12%
2017	25	3598	0.69%
2018	13	3690	0.35%
2019	4	3893	0.10%
Grand Total	428	49263	0.87%

Table 4. The selection of variables from raw financial data.

	Raw Financial Variables Selected in the American Financial Reports and Capital Markets (Bao, 2020)	Corresponding Raw Financial Variables in Chinese Financial Reports and Capital Markets
Balance Sheet	(1) Cash and short-term investments	(1) Cash + Short-term Investments
	(2) Receivables, total	(2) Accounts Receivable + Notes Receivable
	(3) Inventories, total	(3) Inventories
	(4) Short-term investments, total	(4) (Before 2007) Net Value of Short-term Investment; (5) (2007 and beyond) Fair Value Through Other Comprehensive Income
	(5) Current assets, total	(6) Current Assets

Table 4. *Cont.*

	Raw Financial Variables Selected in the American Financial Reports and Capital Markets (Bao, 2020)	Corresponding Raw Financial Variables in Chinese Financial Reports and Capital Markets
	(6) Property, plant and equipment, total	(7) Total Tangible Assets
	(7) Investment and advances, other	(8) (Before 2007) Net Value of Short-term Investment + Net Value of Long-term Investment; (9) (From 2007 to 2017) Investment Property + Trading Financial Assets + Available-for-sale Securities + Held-to-maturity Investments + Long-term Equity Investments (10) (2018) Investment Property + Trading Financial Assets + Other Equity Instrument Investment + Long-term Equity Investments
	(8) Assets, total	(11) Total Assets
	(9) Accounts payable, trade	(12) Accounts Payable
	(10) Debt in current liabilities, total	(13) Notes Payable + Non-current Liabilities maturing within one year
	(11) Income taxes payable	(14) Income Tax Expense − Deferred Income Tax Liabilities + Deferred Income Tax Assets
	(12) Current liabilities, total	(15) Total Current Liabilities
	(13) Long-term debt, total	(16) Total Long-term Liabilities
	(14) Liabilities, total	(17) Total Current Liabilities
	(15) Common/ordinary equity, total	(18) Total Owner's Equity − Preferred Stock
	(16) Preferred/preference stock (capital), total	(19) Preferred Stock
	(17) Retained earnings	(20) Surplus Reserve + Retained earnings
Income Statement	(18) Sales/turnover (net)	(21) Operating Income
	(19) Cost of goods sold	(22) Operating Cost

Table 4. *Cont.*

	Raw Financial Variables Selected in the American Financial Reports and Capital Markets (Bao, 2020)	Corresponding Raw Financial Variables in Chinese Financial Reports and Capital Markets
	(20) Depreciation and amortization	(23) Depreciation for Fixed Assets + Amortization of Intangible Assets + Amortization of Long-Term Expenses Prepayments
	(21) Interest and related expenses, total	(24) Selling Expenses + Administration Expenses + Financial Expenses
	(22) Income taxes, total	(25) Income Tax Expense
	(23) Income before extraordinary items	(26) Net Profit + Non-operating Revenue − Non-operating Expenses
	(24) Net income (loss)	(27) Net Profit
Cash Flow Statement	(25) Long-term debt issuance	(28) Bonds Payable
Market Value	(26) Sale of common and preferred stock	(29) Annual Turnover of Tradable Shares
	(27) Price close, annual, fiscal	(30) Annual Closing Price
	(28) Common shares outstanding	(31) Year-end Circulation Market Value

In February 2006, the Ministry of Finance issued 39 new accounting standards, and the listed companies in China were required to implement these accounting standards from 1 January 2007 [43]. Since then, the Ministry of Finance has successively issued three accounting standards for financial instruments. Since the sample covers the period from 1998 to 2016, this study considers the revision of the accounting standards, generates the financial variables involved according to the corresponding calculation formulas for different periods, and keeps the calculation method of the remaining financial variables unchanged.

4.3. Data Preprocessing

There are some missing values in the financial reporting data in the research samples obtained from the CSMAR database. To evaluate the performance of this model in the Chinese market, this study adopts the same approach as that of Bao et al. (2020) and retains the missing values to avoid noise caused by other filling forms. In this study, when the values of 28 characteristic variables in a sample are missing, the missing value is retained. For example, if "Preferred Stock" is null in a sample in 2012, the null value is retained instead of being filled with averages or other values.

Among all samples selected in this paper, the fraud samples involving fictitious profits announced by the CSRC are labeled as 1, and the remaining non-fraud samples are labeled as 0.

4.4. Evaluation Indicators

In the existing research, the detection of financial reporting fraud is often regarded as a typical two-class classification task, so the performance of the financial fraud model can be measured by using the evaluation indicators for classification tasks. The evaluation indicators mainly include the following two categories: The indicators based on the confusion matrix including accuracy, precision, true positive rate (sensitivity), true negative rate (specificity), false positive rate, false negative rate, and F1 score; the indicators based on the ROC curve, including AUC [44,45].

Due to the small proportion of fraud samples, the accuracy cannot properly reflect the model's prediction performance. Considering that the commit of fraud will bring significant harmful effects to the enterprise and the market. This paper selects precision, sensitivity, and AUC as evaluation indicators to explain more critical information. Specifically, precision measures the percentage of all samples predicted to be fraudulent are really fraudulent. Sensitivity measures the percentage of all fraud samples correctly predicted to be fraud samples. The ROC curve is a curve with FPR as the abscissa and TPR as the ordinate. The AUC is equal to the area under the ROC curve. The AUC considers both the accuracy of positive sample identification and the false positive rate of negative samples. Thus, it comprehensively investigates the true positive rate and true negative rate, and its value falls within [0, 1]. The closer the AUC value is to 1, the better the prediction effect of the model.

4.5. Model Construction

As the internal stakeholders of the enterprise will try their best to hide the fraud, external regulators need to spend much time to obtain sufficient evidence, and the verification process will consume a lot of staffing and material resources. Although the companies that are found to have committed fraud will be made public, in practice, due to a certain interval (i.e., the lag period) between the time when the listed companies in China commit fraud and the time when the fraud is found, it is not possible to obtain all the fraud samples in the lag period. To guarantee the generalization performance of the model, the lag period is considered when setting the training set and the testing set.

This paper uses the data of listed companies in China from 1998 to 2009 as the training set, and the data of listed companies in 2011, 2012, 2013, 2014, 2015, and 2016 as the testing set. Specifically, six groups of experiments are conducted. Each group of experiments selects the sample data of not less than 10 years as a training set and uses the sample data in the T+2 year as the test set. For example, if the sample data from 1998 to 2009 are used as the training set, the sample data in 2011 are used as the test set. If the sample data from 1998 to 2010 are used as the training set, the sample data in 2012 are used as the test set, etc.

According to the divided sample data set, this study uses Python language to call the RUSBoostClassifier module, and constructs a financial fraud detection model based on RUSBoost algorithm. Referring to the design parameters of the RUSBoost model, the number of decision trees is set at 3000, the minimum number of leaf samples is 5, the learning rate is 0.1, and the ratio of the number of fraud samples to the number of non-fraud samples is 1:1.

The RUSBoost algorithm is a common ensemble learning algorithm for unbalanced samples. It forms a new balanced sample data set by randomly under-sampling a certain proportion from a majority of data sets and then combining it with a minority of samples for unbalanced data sets.

4.6. Model Results

According to the evaluation indicators selected in this paper, the results are shown in Table 5.

Table 5. The results of the RUSBoost model.

Training Set Period	Testing Set Period	AUC	Sensitivity	Precision
1998~2009	2011	0.636	47.4%	61.4%
1998~2010	2012	0.656	51.8%	62.8%
1998~2011	2013	0.597	56.6%	67.5%
1998~2012	2014	0.586	52.9%	60.8%
1998~2013	2015	0.634	48.5%	61.5%
1998~2014	2016	0.493	54.1%	63.8%
	Average Value	0.600	51.9%	63.0%
	Bao's model value	0.725		

According to the model's results, the average AUC is 0.60, which is far lower than that of the model proposed by Bao et al. (with an AUC of 0.725). Thus, Hypothesis 1 is denied. The RUSBoost model suitable for the Western markets does not achieve the same prediction effect in China. Among them, the average sensitivity of the model is 51.9%, which indicates that among the listed companies in China, the fraud samples judged correctly by the model account for only half of the actual fraud samples, and the identification performance is low. The average precision of the model is 63%, indicating that 63% of the samples that are judged as fraud are fraudulent samples. Though the model can assist regulators in making pre-judgment to a certain extent, the identification ability is far lower than the effect applied to Western markets.

5. Improvement of the Financial Reporting Fraud Detection Model

5.1. Model Optimization

Through the analysis of the results, this study considers that the sample data set needs to be redesigned due to the extreme imbalance of positive and negative samples and the difference between the fraud paradigm in the Chinese market and that in the United States market, which has a significant impact on the detection performance. The following are two aspects of improvement considered in this paper:

1. The extreme imbalance of positive and negative samples.

Financial report fraud detection is a typical extreme imbalance sample problem. As stated in Section 4.1, among the Chinese listed companies, the proportion of companies with fictional profits fluctuates around 1% each year. Compared with non-fraudulent companies, fraudulent companies account for only a small proportion. Under an extreme imbalance of positive and negative samples, the random selection of non-fraudulent samples based on fraudulent samples brings great volatility. Therefore, this paper reduces the scope of non-fraud samples based on the industry, year, and asset size of the fraud samples to reduce the volatility caused by random sampling of non-fraud samples.

This paper selects matched non-fraud samples for each fraud sample under the following restrictions: first, the corresponding sample companies and the fraud sample companies are in the same industry; second, the corresponding sample and the fraud sample are in the same year; and third, the asset size of the corresponding sample ranges from 95% to 105% that of the fraud sample. Based on 337 fraud samples from 1998 to 2016, 4463 non-fraud samples were selected. In this study, six data sets are constructed, with the samples from 1998~2009 as the training set and the samples from 2011~2016 as the test set. In each constructed data set, non-fraud samples are randomly selected and matched with fraud samples. Aiming at the problem of continuous fraud, to avoid exaggerating the prediction effect, for the samples of listed companies with continuous fraud in the training set, this study changes the fraud sample mark "1" to the non-fraud sample mark "0" [10].

2. Adjustment of interval time between the training set and testing set.

When the previous model is built, the interval between the training set and the testing set is 2 years. However, after statistics and calculations, it is found that the average time spent in detecting fraud of listed companies in China in this study is about 4 years. The interval of some samples is presented in Table 6.

Table 6. The sample example of the time interval between fraud year and announcement time.

Securities Code of Sample Company	Year of Fraud Practices	Year of Fraud Announcement	Interval Time
000020	2000	2004	4
000156	2001	2007	5
000405	1998	2002	4
000408	2017	2019	2
000430	1996	2001	5
000508	1996	1998	2
000509	2000	2005	5
000510	1997	2001	4
000511	2015	2017	2
000514	1998	2003	5
000519	2014	2018	4
000529	2003	2007	4

To avoid the influence of a lag period on model performance, this study set the T+4th year of the last year of the training set as the year of the testing set. In the newly constructed six data sets, the samples in 1998–2007 can be taken as the training set and those in 2011 as the testing set; the samples in 1998~2008 can be taken as the training set and those in 2012 as the testing set. Similarly, the average value of the model tests of the six data sets is taken as the test result. The improved data set is shown in Table 7.

Table 7. The division of the training set and test set.

Dataset Number	Training Set Period	Testing Set Period	Interval Time
1	1998~2007	2011	4 years
2	1998~2008	2012	4 years
3	1998~2009	2013	4 years
4	1998~2010	2014	4 years
5	1998~2011	2015	4 years
6	1998~2012	2016	4 years

5.2. The Results of the Improved Model

To construct a fraud detection model suitable for listed companies in China, this paper selects two single classifiers and five ensemble classifiers for comparison. The single classifiers include logistic regression, SVM, and ensemble classifiers involving random forest, AdaBoost, XGBoost, RUSBoost, and stacking. Among them, the stacking algorithm can form a robust classifier by combining multiple weak classifiers. Particularly, the weak classifier of the stacking algorithm can be a single classifier or an ensemble learning algorithm. The stacking algorithm used in this paper is ensembled by six classifier algorithms, including numbers 1–6 in Table 8. The evaluation indicators are AUC, sensitivity, and precision.

Table 8. The statistics of the improved model results.

Number	Model	AUC	Sensitivity	Precision
	RUSBoost (Before readjustment)	0.600	51.9%	63.0%
1	RUSBoost	0.721	64.7%	64.7%
2	Logistic Regression	0.697	58.8%	61.7%
3	SVM	0.679	70.6%	67.6%
4	AdaBoost	0.723	64.7%	70.6%
5	Stacking	0.742	76.5%	76.5%
6	Random Forest	0.739	70.6%	70.6%
7	XGBoost	0.740	70.6%	73.5%

Based on Python language, this study uses different modules for each algorithm functions. For single classifier, the study uses the LogisticRegression module and the svm_SVC module. For ensemble classifiers, the study uses the RUSBoostClassifier module, the RandomForestClassifier module, the AdaBoostClassifier module, the XGBClassifier module and the StackingClassifier module. This study evaluates the performance of the model through evaluation indicators such as AUC, sensitivity, and precision. The model results are shown in Table 8.

The results show that after the re-adjustment mentioned above of the sample data set of listed companies in China, the AUC of the fraud detection model based on the RUSBoost algorithm rises to 0.721, the sensitivity rises from 0.519 to 0.647, and the precision rate rises from 0.63 to 0.647. The change in sensitivity shows that the proportion of samples successfully detected by the model in the actual fraud samples has increased from 51.9% to 64.7%, indicating that the ability of the model to detect fraud samples has been significantly improved. Note that this study compares the detection performance of single classifiers and ensemble classifiers on the unified sample data set. The AUC of RUSBoost, random forest, AdaBoost, XGBoost, and stacking reaches 0.721, 0.739, 0.723, 0.740, and 0.742, while the AUC of logistic regression and SVM reaches 0.697 and 0.679, respectively. Therefore, the AUC of the ensemble classifier is generally larger than that of the single classifier. Considering that AUC comprehensively investigates the true positive rate and true negative rate, the comparison results of AUC indicate that the overall performance of the ensemble classifier is better than that of the single classifier, which supports Hypothesis 2. In addition, compared with the AUC value of the original RUSBoost model (i.e., 0.725), random forest, XGBoost, and stacking can achieve better detection performance for financial fraud report identification of listed companies in China.

Among them, the stacking model has achieved better performance. From the sensitivity index, the sensitivity of the stacking model reaches 76.5%, indicating that the fraud samples correctly judged by the model account for 76.5% of the actual fraud samples. Thus, the model can identify most fraud samples and has specific practical application value. From the precision index, the precision of the stacking model can reach 76.5%, indicating that the samples that are judged to be fraudulent account for 76.5% of actually fraudulent samples, and the judgment efficiency is relatively high.

Each test system will randomly select non-fraud samples equal to the number of fraud samples from the sample data set, which will bring certain fluctuations to each result. To evaluate the stability of each model, this study repeats the selection process 20 times to statistically analyze each result, and the standard deviation is calculated. The statistical results of stability are shown in Table 9.

Table 9. The standard deviation of the results of the improved model.

	Dataset Number						
Model	**1**	**2**	**3**	**4**	**5**	**6**	**Average Value**
RUSBoost (before readjustment)	0.057	0.053	0.062	0.032	0.037	0.051	0.049
RUSBoost	0.066	0.072	0.064	0.065	0.060	0.051	0.063
Logistic Regression	0.076	0.047	0.052	0.048	0.047	0.047	0.053
SVM	0.050	0.062	0.066	0.057	0.034	0.052	0.054
AdaBoost	0.071	0.071	0.065	0.035	0.047	0.050	0.057
Stacking	0.043	0.054	0.056	0.060	0.041	0.053	0.051
Random forest	0.066	0.074	0.048	0.047	0.049	0.041	0.054
XGBoost	0.078	0.063	0.055	0.051	0.043	0.036	0.054

The results indicate that, compared with other algorithms, the stacking model has better stability and can keep the detection performance consistent under multiple tests.

By comparing various machine learning methods, this study finds that the AUC of the stacking model can reach 0.742, indicating that the model is effective for detecting financial reporting fraud of listed companies in China and can keep the detection performance relatively stable. The financial reporting fraud detection model of listed companies in China based on the stacking algorithm proposed in this paper can predict whether financial reporting fraud has occurred in listed companies in practice, thus helping supervisors, investors, and others to focus on listed companies that are predicted to be fraudulent. With the help of the model, the labor and time costs in the preliminary investigation can be significantly reduced, and the judgment efficiency of investors and regulators can be greatly improved.

6. Conclusions and Implication

This study makes an empirical study on the detection of financial fraud based on machine learning model. Based on the background of China's capital market, the study selects Chinese listed companies from 1998 to 2016, and selects 28 groups of raw data from financial reports. For financial reporting of fraud of Chinese listed companies, the study constructs machine learning models based on two single-classification machine learning algorithms and five ensemble learning algorithms, and compares the performance of each model. The research results indicate that the ensemble classifier generally has a better detection effect than the single classifier. AdaBooat, XGBoost, RUSBoost, random forest and stacking can effectively detect fraud of listed companies in China. Among them, a financial reporting fraud detection model suitable for listed companies in China can be better constructed based on the stacking algorithm, with an overall AUC of 0.742 and sensitivity and precision of 76.5%, which can effectively detect financial reporting fraud.

There are certain theoretical and practical implications of the research findings presented in this paper. From the perspective of theoretical implications, this paper extracts raw data from financial reports to provide a new perspective to quickly detect fraud in Chinese listed companies from financial reports. This is also the first time to propose a stacking algorithm-based financial reporting fraud identification model for listed companies in China, which enriches the application of the machine learning method in financial reporting fraud identification. Meanwhile, for financial report fraud detection, this study finds that the ensemble classifier has better detection performance than the single classifier, which can provide a specific reference for choosing the machine learning method in other similar scenarios. Besides, there are also certain practical implications of this study. Firstly, the study proposes that the raw financial report data can be directly used for fraud detection in China, which reduces the work of data processing and the practical difficulty of fraud detection. Secondly, this study uses the stacking algorithm to construct a new, effective financial reporting fraud identification model for listed companies in China, which provides a simple and effective approach for investors, regulators, and management.

Author Contributions: Conceptualization, Z.W.; Methodology, Z.W.; Validation, Y.C.; Writing—review & editing, Y.C.; Supervision, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, grant number 72172132.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. 2020 Global Study on Occupational Fraud and Abuse. Available online: https://legacy.acfe.com/report-to-the-nations/2020/ (accessed on 2 November 2022).
2. Jia, C.; Ding, S.; Li, Y.; Wu, Z. Fraud, enforcement action, and the role of corporate governance: Evidence from China. *J. Bus. Ethics* **2009**, *90*, 561–576. [CrossRef]
3. Mitchell, T.M. *The Discipline of Machine Learning*; Carnegie Mellon University, School of Computer Science, Machine Learning Department: Pittsburgh, PN, USA, 2006.
4. Perols, J. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Audit. A J. Pract. Theory* **2011**, *30*, 19–50. [CrossRef]
5. Song, X.-P.; Hu, Z.-H.; Du, J.-G.; Sheng, Z.-H. Application of machine learning methods to risk assessment of financial statement fraud: Evidence from China. *J. Forecast.* **2014**, *33*, 611–626.
6. Dietterich, T.G. Ensemble Methods in Machine Learning. In *International Workshop on Multiple Classifier Systems*; Springer: Berlin/Heidelberg, Germany, 2000; pp. 1–15.
7. Ballings, M.; Van den Poel, D.; Hespeels, N.; Gryp, R. Evaluating multiple classifiers for stock price direction prediction. *Expert Syst. Appl.* **2015**, *42*, 7046–7056. [CrossRef]
8. Duin, R.P.W. The combining classifier: To train or not to train? In *Object Recognition Supported by User Interaction for Service Robots*; IEEE: Piscataway, NJ, USA, 2002; Volume 2, pp. 765–770.
9. Dechow, P.M.; Ge, W.; Larson, C.R.; Sloan, R.G. Predicting material accounting misstatements. *Contemp. Account. Res.* **2011**, *28*, 17–82. [CrossRef]
10. Bao, Y.; Ke, B.; Li, B.; Yu, Y.J.; Zhang, J. Detecting accounting fraud in publicly traded US firms using a machine learning approach. *J. Account. Res.* **2020**, *58*, 199–235. [CrossRef]
11. Sun, Y.; Ma, Z.; Zeng, X.; Guo, Y. A Predicting Model for Accounting Fraud Based on Ensemble Learning. In Proceedings of the 2021 IEEE 19th International Conference on Industrial Informatics, Palma de Mallorca, Spain, 21–23 July 2021.
12. Tang, J.; Karim, K.E. Financial fraud detection and big data analytics—Implications on auditors' use of fraud brainstorming session. *Manag. Audit. J.* **2018**, *34*, 324–337. [CrossRef]
13. Yao, J.; Zhang, J.; Wang, L. A financial statement fraud detection model based on hybrid data mining methods. In Proceedings of the 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 26–28 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 57–61.
14. Bao, Y.; Hilary, G.; Ke, B. Artificial Intelligence and Fraud Detection. In *Innovative Technology at the Interface of Finance and Operations*; Babich, V., Birge, J.R., Hilary, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2022; Volume 1, pp. 223–247.
15. Skousen, C.J.; Smith, K.R.; Wright, C.J. Detecting and Predicting Financial Statement Fraud: The Effectiveness of the Fraud Triangle and SAS No. 99. In *Corporate Governance and Firm Performance*; Hirschey, M., John, K., Makhija, A.K., Eds.; Emerald Group Publishing Limited: Bingley, UK, 2009; Volume 13, pp. 53–81.
16. Spathis, C.T. Detecting false financial statements using published data: Some evidence from Greece. *Manag. Audit. J.* **2002**, *17*, 179–191. [CrossRef]
17. Kirkos, E.; Spathis, C.; Manolopoulos, Y. Data mining techniques for the detection of fraudulent financial statements. *Expert Syst. Appl.* **2007**, *32*, 995–1003. [CrossRef]
18. Ravisankar, P.; Ravi, V.; Rao, G.R.; Bose, I. Detection of financial statement fraud and feature selection using data mining techniques. *Decis. Support Syst.* **2011**, *50*, 491–500. [CrossRef]
19. Kanapickienė, R.; Grundienė, Ž. The model of fraud detection in financial statements by means of financial ratios. *Procedia Soc. Behav. Sci.* **2015**, *213*, 321–327. [CrossRef]
20. Cecchini, M.; Aytug, H.; Koehler, G.J.; Pathak, P. Detecting management fraud in public companies. *Manag. Sci.* **2010**, *56*, 1146–1160. [CrossRef]
21. Purda, L.; Skillicorn, D. Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. *Contemp. Account. Res.* **2015**, *32*, 1193–1223. [CrossRef]

22. Liu, C.; Chan, Y.; Alam Kazmi, S.H.; Fu, H. Financial fraud detection model: Based on random forest. *Int. J. Econ. Financ.* **2015**, *7*, 178–188. [CrossRef]
23. Ngai, E.W.T.; Hu, Y.; Wong, Y.H.; Chen, Y.; Sun, X. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decis. Support Syst.* **2011**, *50*, 559–569. [CrossRef]
24. Kotsiantis, S.; Koumanakos, E.; Tzelepis, D.; Tampakas, V. Forecasting fraudulent financial statements using data mining. *Int. J. Comput. Intell.* **2006**, *3*, 104–110.
25. Hajek, P.; Henriques, R. Mining corporate annual reports for intelligent detection of financial statement fraud—A comparative study of machine learning methods. *Knowl. Based Syst.* **2017**, *128*, 139–152. [CrossRef]
26. Baesens, B.; Höppner, S.; Verdonck, T. Data engineering for fraud detection. *Decis. Support Syst.* **2021**, *150*, 113492. [CrossRef]
27. Peng, C.Y.J.; Lee, K.L.; Ingersoll, G.M. An introduction to logistic regression analysis and reporting. *J. Educ. Res.* **2002**, *96*, 13–14. [CrossRef]
28. Bewick, V.; Cheek, L.; Ball, J. Statistics review 14: Logistic regression. *Crit. Care* **2005**, *9*, 1–7. [CrossRef]
29. Jakkula, V. Tutorial on support vector machine (svm). *Sch. EECS Wash. State Univ.* **2006**, *37*, 3.
30. Chen, Y.W.; Lin, C.J. Combining SVMs with Various Feature Selection Strategies. In *Feature Extraction*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 315–324.
31. Pavlyshenko, B. Using stacking approaches for machine learning models. In Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, 21–25 August 2018; pp. 255–258.
32. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
33. Biau, G.; Scornet, E. A random forest guided tour. *Test* **2016**, *25*, 197–227. [CrossRef]
34. Wu, X.; Kumar, V.; Ross Quinlan, J.; Yang, Q. Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **2008**, *14*, 1–37. [CrossRef]
35. Schapire, R.E. Explaining Adaboost. In *Empirical Inference*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 37–52.
36. Brownlee, J. *XGBoost with Python: Gradient Boosted Trees with XGBoost and Scikit-Learn*; Machine Learning Mastery: San Francisco, CA, USA, 2016.
37. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
38. Seiffert, C.; Khoshgoftaar, T.M.; Van Hulse, J.; Napolitano, A. RUSBoost: Improving classification performance when training data is skewed. In Proceedings of the 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 1–4.
39. Seiffert, C.; Khoshgoftaar, T.M.; Van Hulse, J.; Napolitano, A. RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **2009**, *40*, 185–197. [CrossRef]
40. Ting, K.M.; Witten, I.H. Stacking Bagged and Dagged Models. In Proceedings of the International Conference on Machine Learning, Nashville, TN, USA, 8–12 July 1997.
41. Sill, J.; Takács, G.; Mackey, L.; Lin, D. Feature-weighted linear stacking. *arXiv* **2009**, arXiv:0911.0460.
42. Huang, S.; Ye, Q.; Xu, S.; Ye, F. Analysis of Financial Fraud of Listed Companies in China from 2010 to 2019. *Financ. Account. Mon.* **2020**, *14*, 153–160. (In Chinese)
43. Wang, M. The Study of Accounting Conservatism after the Promulgation of New Accounting Standards for Enterprises-Based on the Evidence in China's Capital Market. *Int. Bus. Res.* **2013**, *6*, 183. [CrossRef]
44. Elmrabit, N.; Zhou, F.; Li, F.; Zhou, H. Evaluation of machine learning algorithms for anomaly detection. In Proceedings of the 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), Dublin, Ireland, 15–19 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–8.
45. Grandini, M.; Bagli, E.; Visani, G. Metrics for multi-class classification: An overview. *arXiv* **2020**, arXiv:2008.05756.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

International Journal of
Financial Studies

MDPI

Article

Using Deep Reinforcement Learning with Hierarchical Risk Parity for Portfolio Optimization

Adrian Millea * and Abbas Edalat

Department of Computing, Faculty of Engineering, Imperial College London, London SW7 2AZ, UK
* Correspondence: a.millea14@imperial.ac.uk

Abstract: We devise a hierarchical decision-making architecture for portfolio optimization on multiple markets. At the highest level a Deep Reinforcement Learning (DRL) agent selects among a number of discrete actions, representing low-level agents. For the low-level agents, we use a set of Hierarchical Risk Parity (HRP) and Hierarchical Equal Risk Contribution (HERC) models with different hyperparameters, which all run in parallel, off-market (in a simulation). The information on which the DRL agent decides which of the low-level agents should act next is constituted by the stacking of the recent performances of all agents. Thus, the modelling resembles a statefull, non-stationary, multi-arm bandit, where the performance of the individual arms changes with time and is assumed to be dependent on the recent history. We perform experiments on the cryptocurrency market (117 assets), on the stock market (46 assets) and on the foreign exchange market (28 pairs) showing the excellent robustness and performance of the overall system. Moreover, we eliminate the need for retraining and are able to deal with large testing sets successfully.

Keywords: Deep Reinforcement Learning; Hierarchical Risk Parity; Hierarchical Equal Risk Contribution; portfolio optimization; cryptocurrencies; stocks; foreign exchange

1. Introduction

Since Modern Portfolio Theory Markowitz (1952), there have been numerous approaches to portfolio optimization. For an overview see Kolm et al. (2014).

The general problem can be stated simply as follows: at each time point t, we have a set of N assets which increase or decrease in value in the next period $t + \tau$, where τ is generally referred to as the holding period or rebalancing period and is the number of steps in which no trades are made. The value is measured in some risk-free asset, such as cash (or in the case of cryptocurrencies, some stablecoin such as USDT). The goal is to find a partitioning of the whole set of resources available at each step, such that the overall portfolio value (its equivalent in cash/USDT) increases over time.

The sampling frequency of the overall time-series, i.e., the selection of the time-step, if it represents an hour, a day or a minute, is generally chosen depending on the market type, available resources and intuition of the researcher. We use 1 h, i.e., one hour, for all the market types we use in our experiments (cryptocurrencies, stocks, foreign exchange) as this seems to be a good compromise between information content and computational resources needed to learn from such a series. Data at higher granularity (less than 1 h) is harder to obtain (usually behind a paywall) as well, and the volatility is lower at higher granularity, which means that generally, the overall profits will be lower for the same number of points.

In this work we combine two sets of machine learning techniques for portfolio optimization:

- The first, from the financial community, has received great praise and has been used successfully in recent years: Hierarchical Risk Parity (HRP) De Prado (2016) and a

variation of it, Hierarchical Equal Risk Contribution (HERC) Raffinot (2018). We use the acronym HC hereon to refer to both of these hierarchical approaches.

- The second one is one of the most advanced machine learning techniques available today: Deep Reinforcement Learning Mnih et al. (2015). It has been used successfully in many domains, from video games to robotics, and has been shown able to learn complex tasks from scratch, without any prior knowledge of the environment. For more details on DRL and its general applications see Li (2017), for DRL in finance see Millea (2021); Mosavi et al. (2020).

DRL provides flexibility and adaptability at the expense of extended computation time and large variance whereas the HC approaches provide some reliability, smaller variance and require less computation time.

One of the major disadvantages of the DRL approach when dealing with the portfolio optimization problem is the continuous multi-dimensional action space needed to model the portfolio allocation weights, which is notoriously hard to explore. Since many iterations are needed for a DRL algorithm to converge, with the duration of each iteration being directly dependent on the state space dimensionality, it is important that the historical time window and the number of assets fed in be reasonably small, as this would increase the state space's dimensionality.

The HC approach on the other hand is very fast but quite limited and strongly influenced by the covariance estimation period. Some HC models can produce very different results depending on this period.

In our work we harness the best of each approach in the following sense: (i) we use a DRL agent for high level decision-making, with a discrete action space, thus overcoming the continuous multi-dimensional action space exploration issue and (ii) avoid the large variance by using as low-level agents HC models with different covariance estimation periods which interact with the environment. Thus, the high-level DRL agent selects among a set of low-level agents which in turn are fast HC models which give a reasonable performance. We could say we are combining weak learners (HC models) with DRL to obtain a strong learner (the overall system).

By combining these two techniques in this way, the agent is able to adapt to changing market regimes, being able to deal well with bear but also bull sessions, thus enabling a much larger testing period without any retraining. It actually removes the need for retraining altogether.

Our contributions are fourfold:

- We combine DRL with HRP and HERC Raffinot (2018) (together referred to as HC—hierarchical clustering) models, to manage the risk and combine the capabilities of different models adaptively based on the current market conditions.
- We use HRP and HERC models with much lower covariance estimation periods than used on other markets or than initially designed for Burggraf (2021); De Prado (2016). We use periods as low as 30 and up to 700 h for covariance estimation. These numbers are usually days or even months in the related literature.
- The whole system enables much larger periods for testing, without any retraining. In the literature they usually employ retraining after short periods of testing. In our case the testing periods (or out-of-sample as they are called in the financial community) are as long as the training period or even longer.
- The network architecture and engineering is minimal, as we use a simple feed-forward network with two hidden layers. This makes the whole system very easy to understand and holds the promise for significant enhancements.

This works as a switching system, which allows the HC models to be switched between each (high-level) time-step. This is useful especially for cryptocurrency markets, where conditions are continuously changing, which HC models are not always able to capture.

Moreover, we are able to break the testing-after-training requirement, and need for retraining, which is present throughout the financial community using DRL (Hirsa et al.

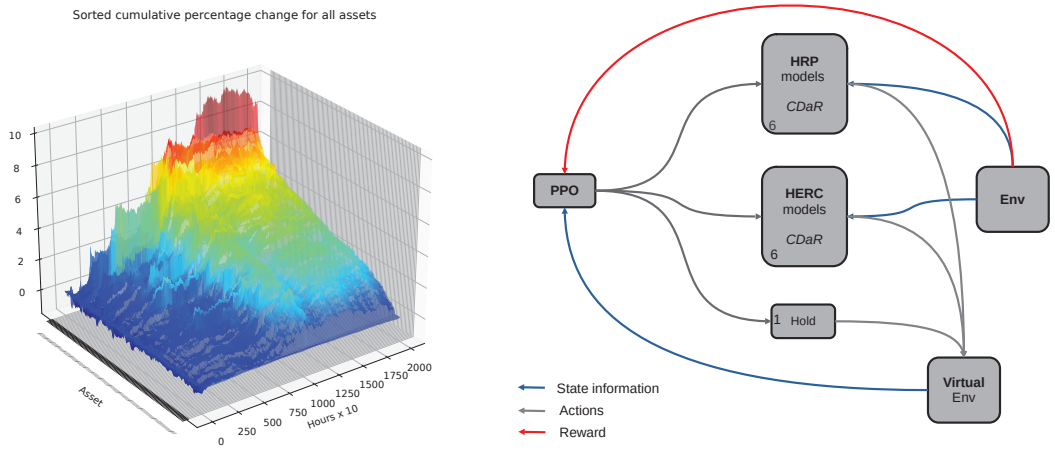
2021; Jiang and Liang 2017; Yang et al. 2020) and use much larger testing periods than generally seen in the literature.

We show in Figure 1a what the cumulative rate of change is for all used crypto. We first compute the normalized return $\frac{p_t}{p_{t-1}} - 1$ (where p_t is the price at time t) at time t and then add these all up for each asset, to get a cumulative indication of growth. We then sort by the last value for cleaner visualisation.

In Figure 1b we show the overall architecture of our system. The integers to the bottom left of the central rectangles denote the number of different models of each type available as actions. Thus we have six HRP models, six HERC models and an additional action for holding. This means that no model is selected and no trade is made.

Having multiple decision-making levels opens up a new class of trading algorithms, flexibly combining specialised agents (this specificity can come from multiple sources, e.g., can be risk focused or dataset driven) which work best in their representative scenarios, based on other criteria, to maximise the overall profits. Numerous applications are possible. In principle any asset allocation problem can be modelled in this way, assuming the existence of good-enough low-level agents (weak learners).

This demonstrates a means of avoiding the cumbersome high-dimensional continuous action space problem. Moreover, it scales well when adding other asset classes (assuming each new class will need at least a few different base-models), adding only one discrete action to the DRL agent for each new model added. In short, we can scale our system to a very large number of assets (assuming we can partition them pertinently into different asset classes) with minimum overhead for the DRL agent.



(**a**) (**b**)

Figure 1. Overview of the data (**left**) and system architecture (**right**). (**a**) Cumulative rate of change of all assets in the portfolio; (**b**) Overview of the system architecture.

2. Related Work

An asset allocation problem tries to optimize some relation between returns and risk. The original efficient frontier, devised by Markowitz is found by maximizing returns and minimizing risk (the risk is generally considered to be the volatility of the portfolio). This is given by the standard deviation of the random variable $w^T X$ where w is the vector of asset weights and X is the matrix of asset returns. Thus the volatility is given by: $\sigma(w) = \sqrt{w^T \Sigma w}$, with Σ the covariance of the asset returns. There are many related measures, among which we mention the ubiquitous ones:

- **minimum variance** portfolio Bodnar et al. (2017); Clarke et al. (2011) aims at minimizing the variance by finding the minimum weight vector w to minimize $w^T \Sigma w$
- **maximum diversification** Choueifaty and Coignard (2008) defines a diversification ratio $D = \frac{w^T \sigma}{\sqrt{w^T \Sigma w}}$

- **the risk parity** solution Maillard et al. (2010) aims at solving a fixed point problem: $w_i = \frac{\sigma(w)^2}{(\Sigma w)_i N}$ with N the number of assets. An alternative formulation is to find w which minimizes
$$\sum_{i=1}^{N}\left[w_i - \frac{\sigma(w)^2}{(\Sigma w)_i N}\right]^2.$$
- **hierarchical clustering approaches** in a series of relatively recent papers. Several works have leveraged the idea of hierarchical clustering which we describe in more detail below.

The initial Critical Line Algorithm (CLA) devised by Markowitz is a quadratic optimization problem with inequality constraints. The CLA guarantees finding the solution in a specific number of iterations. It also bypasses the Karush–Kuhn–Tucker conditions admirably. Even though it is a remarkable algorithm is has some downsides:

- (i) if the return predictions deviate by even a small amount from the actual returns then the allocation changes drastically Michaud and Michaud (2007);
- (ii) it requires the inversion of the covariance-matrix, which can be ill-conditioned and thus the inversion is susceptible to numerical errors;
- (iii) correlated investments indicate the need for diversification but this correlation is exactly what makes the solutions unstable (this is known as the Markowitz curse).

For more details, we refer the curious reader to De Prado (2016). Some solutions were developed to deal with the above issues. In one solution, the prediction of the returns was dropped altogether, which leads to risk parity approaches Jurczenko (2015). Numerical stability of the matrix inversion procedure has also been improved through other techniques such as shrinking Ledoit and Wolf (2003). For more details about the instabilities mentioned above see Kolm et al. (2014). Other approaches have been explored in the literature, such as incorporating prior beliefs Black and Litterman (1992) and even online reinforcement learning was used Moody and Saffell (1998).

More recently, a new set of simple and elegant techniques was introduced De Prado (2016), where clustering is employed, unveiling a hierarchical structure in the correlations and allowing resources to be allocated depending on this hierarchy. In this case, there is no need for a matrix inversion, thus avoiding the instabilities associated with it. Following this initial algorithm, other techniques making use of hierarchical clustering have been developed (HERC Raffinot (2018), HCAA Raffinot (2017)).

2.1. Hierarchical Risk Parity

Next, we will describe in some detail the initial algorithm given in De Prado (2016) and will show at each step the visualisation corresponding to our own data on the cryptocurrency market. Given a set of N assets, each of length T, we have a $N \times T$ matrix of assets. There are three main steps:

- 1. Computing the dendrogram using hierarchical tree clustering algorithm by using a distance matrix of the asset correlations.
  - First we compute the correlation among assets, which produces an $N \times N$ matrix we call ρ.
  - Then, we convert this to a distance matrix through
$$D(i,j) = \sqrt{0.5 \times (1 - \rho(i,j))}.$$
  - We next devise a new distance matrix $\widehat{D}$ which is actually a similarity matrix. It is computed as follows:
$$\widehat{D}(i,j) = \sqrt{\sum_{k=1}^{N}(D(k,i) - D(k,j))^2}.$$

– We then start to form clusters as follows:

$$U = argmin_{i,j}\widehat{D}(i,j).$$

This gives us two indices r_1 and c_1, which are the row and column indices in $\widehat{D}$. We combine the assets given by the respective indices to obtain the first cluster $U[1]$. For more details on the procedure please see the original paper De Prado (2016).

– We now delete the newly formed cluster assets row and column from the matrix (r_1 and c_1) and replace them with a single one. Computing distances between this new cluster and the remaining assets can be performed by multiple different criteria, as is the general case with clustering, either single-linkage, average-linkage, etc.

– We repeat the last two steps until we have only one cluster. So we can see this is an agglomerative clustering algorithm. We show its dendrogram (graphical depiction of the hierarchical clustering) on our crypto data in Figure 2.

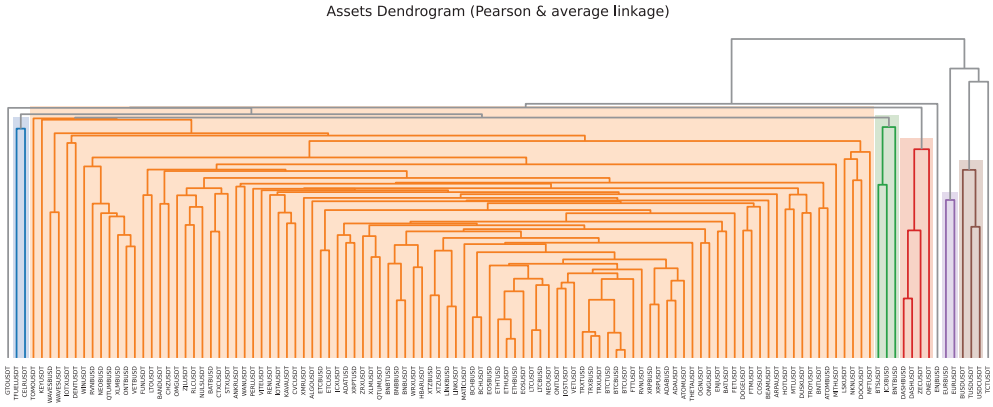


Figure 2. Dendrogram of the hierarchical clustering of the crypto assets.

• 2. Doing matrix seriation. This procedure rearranges data in such a way that larger covariances are closer to the diagonal and that similar values are near. This results in a quasi-diagonal covariance matrix. We show this process on our own crypto dataset Figure 3.

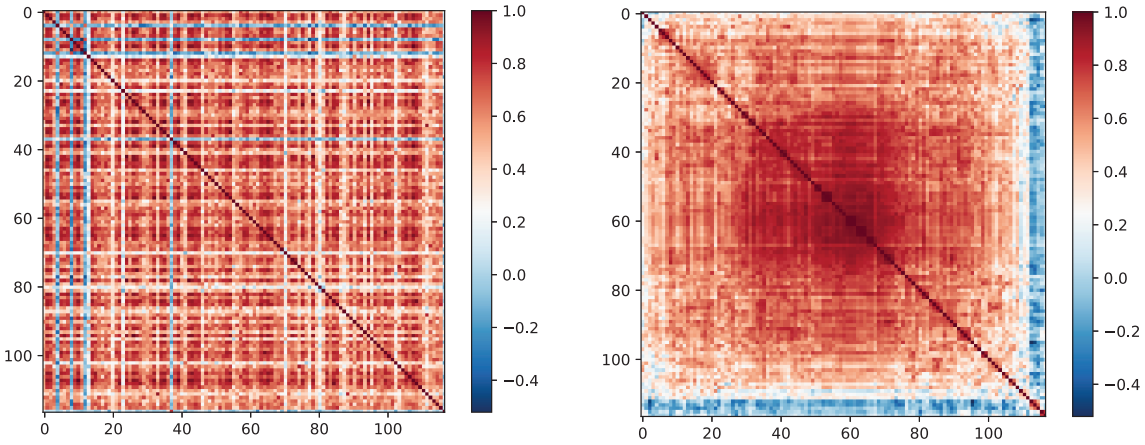


Figure 3. Original correlation distance matrix (**left**) and after matrix seriation or quasi-diagonalization (**right**).

• 3. Recursive bisection. This is the final step which actually assigns weights to the assets using the previous clustering. Looking at the hierarchical tree structure:

– We first set all weights to 1.

– Next, starting from the root, for each cluster we use the following weights to compute the volatilities:

$$w_i = \frac{diag(V_i)^{-1}}{trace(diag(V_i)^{-1})}.$$

This uses the fact that for a diagonal covariance matrix, the inverse-variance allocations are optimal.
– Then for the two branches of each node (a node corresponds to a cluster) we compute the variance: $\sigma_1 = w_1^T V_1 w_1$ and $\sigma_2 = w_2^T V_2 w_2$.
– Finally, we rescale the weights as $w_i = \alpha_i w_i$ with $\alpha_1 = 1 - \frac{\sigma_1}{\sigma_1+\sigma_2}$ and $\alpha_2 = 1 - \alpha_1$

Most HC models evaluated in the literature look at much larger time periods, and are often used for the analysis of the long-term asset behavior. We show that shorter periods can be used successfully and the overall system performance can be enhanced by adaptively selecting among them.

2.2. Deep Reinforcement Learning

In a different line of work, DRL has been applied with some degree of success in many markets Betancourt and Chen (2021); Jiang and Liang (2017); Zejnullahu et al. (2022); Zhang et al. (2020). One of the major differences between the two approaches, is the period at which they are evaluated on, with HRP and related approaches tackling much larger out-of-sample sets. Most DRL algorithms used in trading are generally trained for a long time and then their performance is evaluated for a short test time. This evaluation period needs to come immediately after the training period. The latter is thought to be required due to the nature of the market and the limitations of the agent which cannot generalize to many different dynamics. Another shortcoming of the DRL approach is that it generally requires a continuous action space, which is notoriously hard to explore.

The new DRL techniques applied on the portfolio management problem seemed to be reinventing the wheel, with some approaches even incorporating covariance estimation. Most of them used some of the existing theory but none of the existing practical models. The models coming from the financial community and DRL seemed to be mutually exclusive, despite tackling the same issue. You could either use one or the other. This is where we try to bridge the gap, by employing these more traditional HC models from the financial community, as a starting point and adding the DRL flexibility on top of them, we avoid its large computational requirements, large variance and instability. Our DRL works on a small number of state dimensions, the recent performance of all models. As for actions it selects among available models, 12 + 1 (hold) possible discrete actions.

In our work we overcome these limitations, leveraging a fast (much faster than DRL) and well-received algorithm (HRP) to be able to use a single agent without need for retraining and we use a discrete action space where each action corresponds to selecting one of the HRP and HERC models' weights assignments.

2.3. Model-Based RL

Model-based RL assumes the existence of some abstract simulation of the world, either devised by the engineer, or learned by a machine learning algorithm. This simulation is then used to generate rewards and/or observations for the agent, which allows it to learn from experiences (which are now embedded in the simulation) without interacting with the real environment. Some advantages and constraints of having a model-based RL agent are:

- The interaction with the real environment might be risky, inducing large costs sometimes. Thus having a zero-cost alternative is desirable.
- Having an accurate enough model of the world is necessary for the agent to learn and behave well in the real environment.
- The model is not necessarily a perfect representation of the real world, but it is close enough to be useful.

There are not many published articles using model-based DRL on trading, but there are some: Wei et al. (2019); Yu et al. (2019). However, all of them deal with prediction of the future, and not with the actual trading. We incorporate in our system, a different type of model, a decision-making model, in which we already have a dozen working trading models which we then select from using our DRL agent. Moreover, since the market state is the same independent of the actions taken by the agent, we can compute the performance of many decision-making models in parallel.

3. System Architecture

We use a high-level DRL agent (Proximal Policy Optimization Schulman et al. (2017) —PPO) which selects the next model to act on the market, based on its recent performance.

PPO is a policy optimization DRL algorithm which uses an approximation to a more analytical trust-region policy optimization Schulman et al. (2015). In general it has good properties throughout a range of tasks and reaches some solution in a reasonable amount of time. It is better behaved, in our experience, more stable than other DRL algorithms[1].

We wanted to combine the flexibility of DRL with the efficiency and robustness of HC approaches. The motivation behind this is to harness the strengths of DRL, while overcoming its drawbacks: high-dimensional continuous action space, large state space and large variance in performance. The whole system is designed to adaptively deal with switching market regimes while avoiding the need for retraining, and it does so successfully as seen in the next section.

We first run the low-level models, the 12 HC models on the market to obtain their performance on the full dataset (Figure 4). We see that the performance is worse than the buy-and-hold strategy for most HC models. We save these models' performances to a file (to avoid computing it every time they are requested by the DRL agent in the training phase) and load them up in each RL experiment. These performances of all models will function as state information for the DRL agent, indexed by the step we are at in the dataset. We also tried adding the returns of all assets, or using only the returns as a state but that produced a worse performance than just using the recent performance of all models.

By having multiple HRP and HERC models with different hyperparameters (see the whole set of hyperparameters in Table 1, we vary the covariance estimation period, as it is significantly influencing performance), we ensure that our pool of decision making models is diverse. Thus, we account for different market regimes, and we remove the need for having the testing period right after the training period. This dependency on the recent past is now incorporated in the recent performance of the individual HC models (which is fed as a state to the PPO agent), thus allowing for much larger testing periods. We show good performance on the crypto market for as much as 13 months without any retraining (with training on 13 months as well). For details of the datasets used and train-test splits see Table 2.

The data we chose were the most recent free data we found available at the time of the respective experiments. We tried to split the data similarly even if they had different sizes for each market, trying to keep a large testing set for each, to avoid cherry-picking, which can happen for small testing sets. We considered the fact that we ended up with different periods and different sizes for each market as beneficial for the overall robustness of sampling and evaluation of our system. The crypto data were taken from https://binance.com/ (accessed on 2 June 2022) through the python API, the forex data were taken from https://data.forexsb.com/ (accessed on 29 September 2022), while the stocks data were taken from https://www.alphavantage.co/ (accessed on 5 February 2022). The individual assets were selected based on how much valid data were available for each asset. We chose the assets with the fewest missing points. See the full list of assets in Table 3.

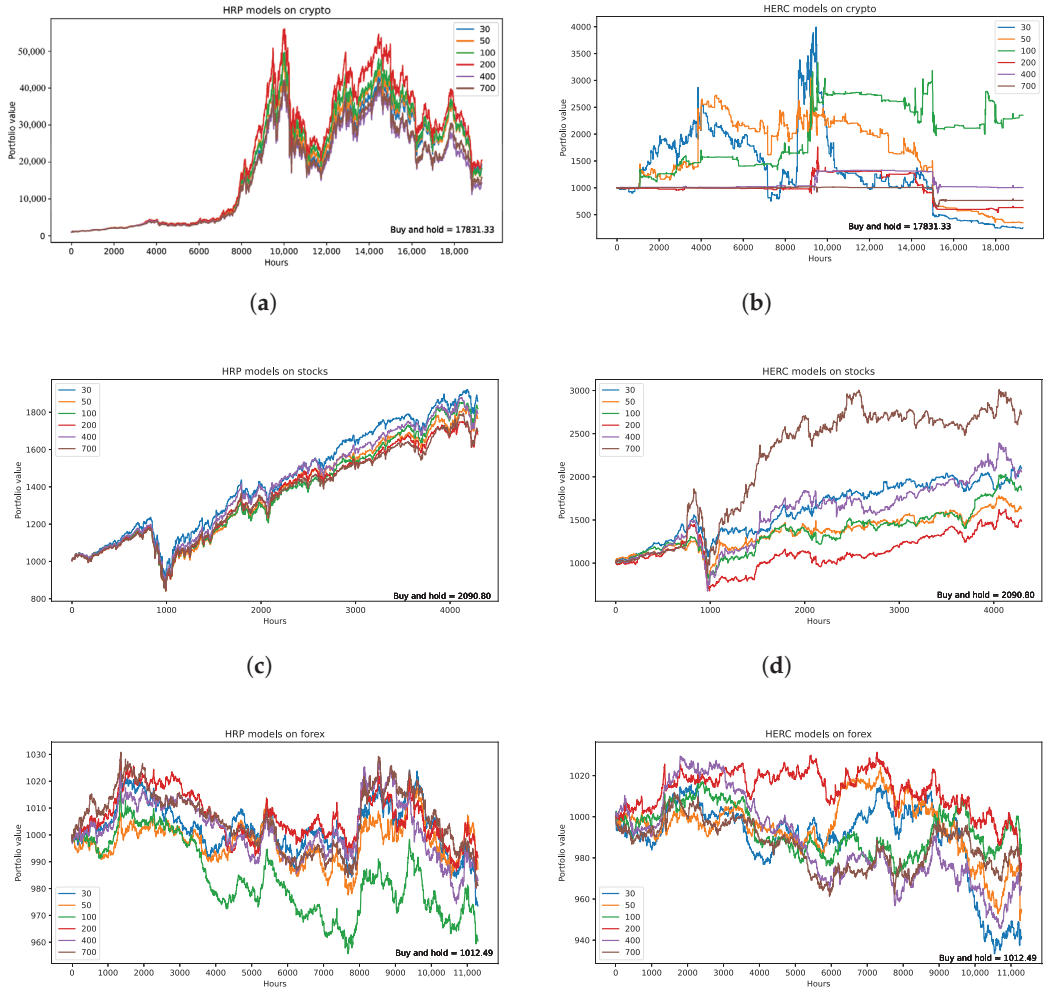
Both the train and test periods start with an offset of 700 h, the maximum covariance estimation period among the HC decision making models. This means that no training or testing takes place in this interval.

Table 1. Hyperparameters for the HRP and HERC models.

Hyperparameter—Model	HRP	HERC
Risk measure	CDaR	CDaR
Codependence	Pearson	Pearson
Linkage	average	average
Max clusters	-	10
Optimal leaf order	True	True

Table 2. Training and testing splits.

Market	Training Set	Testing Set
Crypto	22 March 2020 04:00–13 May 2021 14:00	13 May 2021 15:00–1 June 2022 03:00
Stocks	17 September 2019 15:30–16 December 2020 12:30	16 December 2020 13:30–4 February 2022 15:30
Forex	11 December 2020 01:00–30 September 2021 23:00	10 January 2021 00:00–28 September 2022 06:00



(**a**) (**b**)

(**c**) (**d**)

(**e**) (**f**)

Figure 4. Performance of individual HRP and HERC models. (**a**) HRP on crypto market (117 coins); (**b**) HERC on crypto market (117 coins); (**c**) HRP on stock market (46 stocks); (**d**) HERC on stock market (46 stocks); (**e**) HRP on forex (28 pairs); (**f**) HERC on forex (28 pairs).

Table 3. List of assets used on each market.

Cryptocurrencies		Stocks	Forex
ADABUSD	LINKBUSD	AAPL	AUDCAD
ADATUSD	LINKUSDT	ABT	AUDCHF
ADAUSDT	LSKUSDT	ACN	AUDJPY
ALGOUSDT	LTCBUSD	ADBE	AUDNZD
ANKRUSDT	LTCUSDT	AMD	AUDUSD
ARPAUSDT	LTOUSDT	AMZN	CADCHF
ATOMBUSD	MATICUSDT	ASML	CADJPY
ATOMUSDT	MFTUSDT	AVGO	CHFJPY
BANDUSDT	MITHUSDT	BABA	EURAUD
BATBUSD	MTLUSDT	BAC	EURCAD
BATUSDT	NEOBUSD	BHP	EURCHF
BCHBUSD	NEOUSDT	COST	EURGBP
BCHUSDT	NKNUSDT	CRM	EURJPY
BEAMUSDT	NULSUSDT	CSCO	EURNZD
BNBBUSD	OGNUSDT	CVX	EURUSD
BNBTUSD	OMGUSDT	DIS	GBPAUD
BNBUSDT	ONEUSDT	GOOG	GBPCAD
BNTBUSD	ONGUSDT	HD	GBPCHF
BNTUSDT	ONTBUSD	INTC	GBPJPY
BTCBUSD	ONTUSDT	JNJ	GBPNZD
BTCTUSD	PERLUSDT	JPM	GBPUSD
BTCUSDT	QTUMBUSD	KO	NZDCAD
BTSUSDT	QTUMUSDT	MA	NZDCHF
BUSDUSDT	RENUSDT	MCD	NZDJPY
CELRUSDT	RLCUSDT	MS	NZDUSD
CHZUSDT	RVNBUSD	MSFT	USDCAD
COSUSDT	RVNUSDT	NFLX	USDCHF
CTXCUSDT	STXUSDT	NKE	USDJPY
CVCUSDT	TCTUSDT	NVDA	
DASHBUSD	TFUELUSDT	NVS	
DASHUSDT	THETAUSDT	ORCL	
DENTUSDT	TOMOUSDT	PEP	
DOCKUSDT	TROYUSDT	PFE	
DOGEUSDT	TRXBUSD	PG	
DUSKUSDT	TRXTUSD	PM	
ENJBUSD	TRXUSDT	QCOM	
ENJUSDT	TUSDUSDT	TM	
EOSBUSD	USDCUSDT	TMO	
EOSUSDT	VETBUSD	TSLA	
ETCBUSD	VETUSDT	TSM	
ETCUSDT	VITEUSDT	UNH	
ETHBUSD	WANUSDT	UPS	
ETHTUSD	WAVESBUSD	V	
ETHUSDT	WAVESUSDT	VZ	
EURBUSD	WINUSDT	WMT	
EURUSDT	WRXUSDT	XOM	
FETUSDT	XLMBUSD		
FTMUSDT	XLMUSDT		
FTTUSDT	XMRUSDT		
FUNUSDT	XRPBUSD		
GTOUSDT	XRPTUSD		
HBARUSDT	XRPUSDT		
HOTUSDT	XTZBUSD		
ICXBUSD	XTZUSDT		
ICXUSDT	ZECUSDT		
IOSTUSDT	ZILUSDT		
IOTAUSDT	ZRXUSDT		
IOTXUSDT	KEYUSDT		
KAVAUSDT			

3.1. Performance of Individual HRP Models

We run multiple HRP models with different parameters in parallel since the problem setup allows this without any issues. See the results on the full datasets in Figure 4.

After we receive the prices for all assets for the current timestep, compute each model's respective profits if allowed to trade on the real market. We assume there is some sort of smoothness in model performance, which recent history being indicative of the immediate future, so we are feeding the recent T steps of the model performance (actually its rate $\frac{P_t}{P_{t-1}}$, to make it independent of the value magnitude).

We use the riskfolio library[2] to acquire the HRP and HERC solution, with the Conditional Drawdown at Risk (CDaR) as the risk measure to optimize the portfolio. CDaR is a measure similar to Conditional Variance at Risk (CVar, Uryasev (2000)), but which uses the drawdown instead of the variance of the assets. For more details on this measure see Goldberg and Mahmoud (2017).

In Figure 4 we show the performance of the individual HC models on all three markets. We vary the covariance estimation period, the parameter which influences significantly the portfolio selection at each step. HERC models are less correlated with other HERC models, with HRP models being highly correlated with other HRP models, however the HRP and HERC models are not correlated. The idea is to have more heterogeneous models such that the DRL agent can choose a better performing model when the performance of the current selection goes down.

In Figure 5 we can see the mean performance of the discrete DRL agent on each market (blue) with the black bars representing the sample standard deviation from 30 samples. We report the result of the best configuration of the hyperparameters. See Table 4 for the full list of hyperparameters used for PPO. The orange bars represent the performance of the continuous DRL agent, when we have 12 continuous actions, one value for each low-level agent. This should enable arbitrary policy learning, combining the best of each agent at each step, while still reducing the dimensionality of the action-space significantly. However, even if this showed better than random mean performance and much lower standard deviation (a much desired trait for a DRL agent), the performance was quite poor compared to the discrete agent. The green bars represent the performance of a random high-level policy, where only the DRL agent is replaced with a random policy, while the low-level agents remain the same. This is to directly compare the performance of a DRL policy with a random policy. The red bars represent a base policy which considers investing at the beginning of the evaluation period the same amount of cash (or USDT) in each of the available assets and holding them until the end of the evaluation period.

Table 4. Hyperparameters for the PPO model.

Hyperparameter—Model	**PPO**
Learning rate	grid_search([0.0001, 0.001, 0.01])
Value function clip	grid_search([1, 10, 100])
Gradient clip	grid_search([0.1, 1, 10])
Network	[1024, 512]
Activation function	ReLU
KL coeff	0.2
Clip parameter	0.3
Uses generalized advantage estimation	True

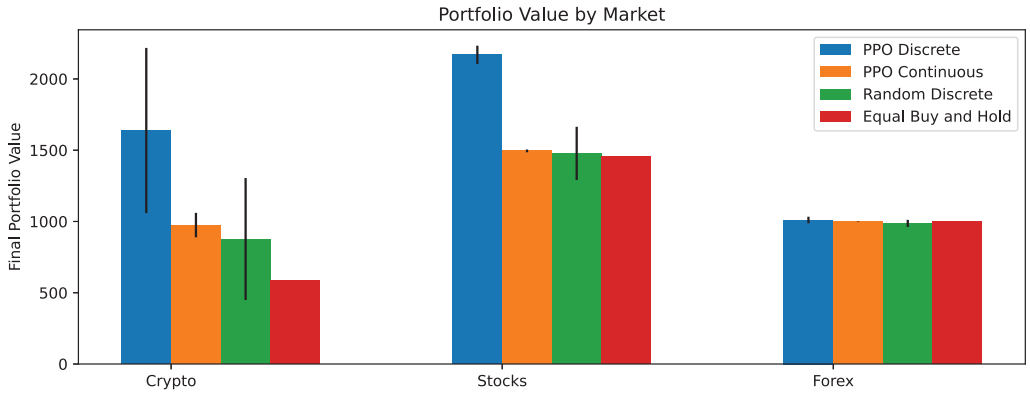


Figure 5. Portfolio value by market.

3.2. Rewards

For RL we used the RLlib library[3]. Aside from the classical reward given to a portfolio optimization agent, we add a hold reward, to incentivize the agent to hold when all the assets are going down. The reward at time t is given by:

$$\text{reward}_t = \sum_{i=1}^{m} q_t^i p_t^i - \sum_{i=1}^{m} q_{t-\tau}^i p_{t-\tau}^i \tag{1}$$

where τ is the number of low-level steps corresponding to a high-level step (we take this as 5), q_t^i is the quantity of asset i at time t, p_t^i is the price of asset i at time t, and q_{t-1}^i is the quantity of asset i at time $t-1$. The quantity of asset i is always positive and is related to the weights w_i^t of the portfolio at time t through the overall previous portfolio value, and is given by:

$$q_t^i = w_t^i \cdot \frac{\sum_{j=1}^{m} q_{t-\tau}^j p_{t-\tau}^j}{p_{t-\tau}^i} \tag{2}$$

The reward for holding is given to the agent only when all assets have gone down and the recent action was hold. This is given by:

$$\text{reward}_{hold} = \sum_{i=1}^{m} \left(1 - \frac{p_{t-\tau}^i}{p_t^i}\right) \tag{3}$$

3.3. Combining the Models

We can also combine the models weighted by some real value (so M continuous outputs, with M the number of models) which is the output of the DRL agent. The previous case is just a simplified version of this, where one of the outputs is 1 and the rest are all 0.

In this manner we can construct arbitrary policies, driven by the DRL agent. We can vary the discretization factor, allowing just a few discrete values for each model. This can be defined as d = *how many values can the weight take between 0 and 1*. If the answer is 2, we are in the previous case, where only one model acts at any point. However, if $d > 2$ then we can have multiple models acting at the same time, with different weights. If d is ∞ then each weight is continuous. We test this version of the system as well, with a continuous-output DDPG Lillicrap et al. (2015) agent in place of the previous discrete-output PPO agent.

Despite training the DDPG agent for 10,000 iterations, it has not learned enough to produce any meaningful results. We omit the results, but we give the formulation here. The poor performance might be because of an insufficient hyperparameter search (see Table 5 for hyperparameters used, and check the RLlib source for the full default set[4]).

Table 5. Hyperparameters for the DDPG model.

Hyperparameter—Model	DDPG
Actor learning rate	grid_search([1×10^{-3}, 1×10^{-4}, 1×10^{-5}])
Critic learning rate	grid_search([1×10^{-3}, 1×10^{-4}, 1×10^{-5}])
Network for actor and critic	[1024, 512]
Activation function	ReLU
Target network update frequency	grid_search([10, 100, 1000])

We also tried a continuous version of the PPO agent, which shows marginally improved performance. Originally PPO was designed as a discrete action space algorithm, but in RLlib it can be used with continuous action spaces as well, by using a Gaussian distribution as the action distribution. We use the same hyperparameters as in the discrete case. Each model is given a weight a_i (with i from 1 to M, in our case $M = 12$) and the portfolio is constructed as follows. First we weigh the asset weights of the respective model by the model weight, which is one element of the output vector, or action, from the DDPG/PPO agent:

$$\bar{\mathbf{w}}_i = a_i \cdot \mathbf{w}_i \tag{4}$$

Then we sum up to weights for each asset, for all models:

$$\bar{w}^j = \sum_{i=1}^{M} \bar{w}_i^j \text{ for } j = 1, \ldots, N \text{ where } N \text{ is the number of assets} \tag{5}$$

Finally, we normalize the weights to sum up to 1:

$$\widehat{w}^j = \frac{\bar{w}^j}{\sum_{j=1}^{N} \bar{w}^j} \tag{6}$$

Normally, if the action components a_i sum up to one then there is no need for the last normalization, however in RLlib they do have strange bounds sometimes.

4. Results

We show good performance on the crypto market for over a year without any retraining and in an extremely bear market. We start with an initial portfolio of USDT/USD 1000 and thus ending for example with USDT 2000 means we have gained 100%.

Since our base models are quite good already, it would be unfair to compare the results of our system (PPO + HC models) with a purely random policy. Thus we replace the PPO agent selecting among the HC models, with a random policy. We use this as a base-model comparison to our PPO agent.

We see in Figure 5 that the variances of both the PPO and random policies are similar for the crypto market (30 samples for each), and that the PPO agent clearly performs better than the random policy on all three markets.

We can attribute the large performance variance due to the fact that we are making decisions at a higher level, among models, and for multiple timesteps. The second baseline in Figure 5 is the buy-and-hold strategy for all assets. The strategy assumes an equal weight attribution for all assets. We see that a random policy at the high-level performs much better than the equal buy-and-hold strategy, at least for the crypto market.

We also see the different magnitudes of gains on different markets, with forex being the least profitable and having the smallest standard deviation. That is due to the nature of the time-series which has very low volatility compared to the crypto and stock market.

The performance is also clearly better than the individual HC models, which shows that our PPO agent managed to learn when to switch between models.

We chose such a large testing set to include both types of market conditions, bull and bear. However, to make sure we deal with widely different market conditions, we switch the training and testing sets. By doing so, we clearly brake the testing-after-training condition usually seen in the community. We show in Figure 6 the results on the stock and crypto markets for our system when the testing set is chronologically before the training set. In this case the crypto market has a very strong bull trend, but our system manages to obtain a significantly higher mean performance (30× vs. 50×). The y axis is on a logarithmic scale for better visualisation. In the stock market our system performs significantly better, even though the market trend is not strongly bullish.

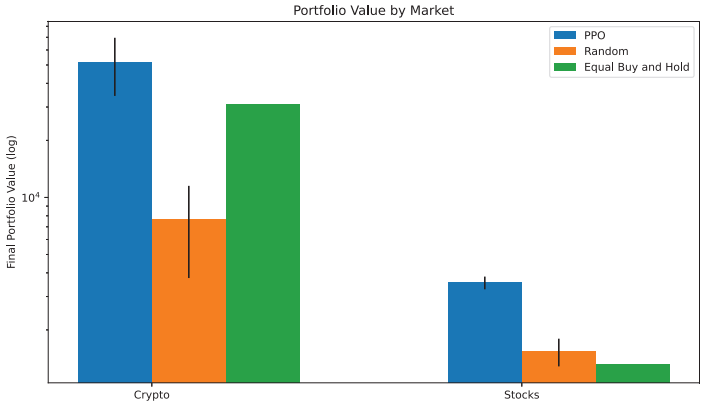


Figure 6. Portfolio value by market.

5. Discussion

From evaluating our system without any retraining on large testing sets, which go through different regimes of the market, we can reliably say that our system works in more market conditions. Even seminal works such as Deng et al. (2016); Jiang et al. (2017) use a few very small isolated testing sets (50 days) with retraining in between. Some do use larger testing sets of 1–2 years, but with larger training sets (3 and up to 9 times) as well Li et al. (2019); Théate and Ernst (2021). Only in Zhang et al. (2020) we see truly large testing sets with a similar size as the training set, but on a different set of assets (a mix of futures), and they still use retraining. We also used different markets and different dataset sizes, adding to the robustness of the evaluation.

We thus add to the body of evidence that the market switches between different dynamics/regimes and having an explicit adaptive mechanism which deals with this switching is highly fruitful. A system with more decision-making layers can include additional useful information at selected layers, thus enabling structured processing of news data or macroeconomic indicators in a more meaningful way. The Ukraine–Russia conflict and the pandemic are such examples of macro-events which would clearly benefit a trading system if incorporated soundly.

6. Conclusions

We combined two of the newest and most performant ML techniques in a hierarchical decision-making system, with a DRL agent working as a high-level agent and selecting between a set of heterogeneous low-level HC agents to effectively act on the trading environment, i.e., assign weights to a portfolio of assets.

We showed that the DRL agent is able to learn when to switch between the low-level models, and that the performance is better than the individual models. We also showed that the performance is better than a random policy, which is a good baseline for the performance of the high level of decision making since the low levels supposedly have reasonable performance, thus a random high-level agent will surely perform better than a purely random agent.

Moreover, we were able to avoid retraining altogether, and increase the size of the testing set significantly, making it as large as the training set. All our results were obtained by using a default neural network with just two layers.

We also attempted to use a continuous action space for the DRL agent and combined ideal weights from each HC model, but this did not work well, or we were not able to find a good hyperparameter configuration for the DDPG agent.

Future work should try to find better hyperparameters, including the architecture of the network. One natural way to extend the current system is to add another level of decision-making and data aggregation, to deal with multiple markets in the same time. More data sources should also be useful, such as social media sentiments, and could be incorporated easily into the system, at selected levels.

The downside of an increasingly hierarchical system is the number of hyperparameters one has to choose from. The more individual agents, the more parameters we have, in addition to the ones used for their interaction. Joint training is prohibitive with a large number of levels, thus a mechanism for pretraining at each level is essential.

However, a hierarchical architecture shows the promise of having specialized and heterogeneous agents which can be selected alternatively to improve the overall system performance. Thus, a more thorough search for base models should be useful, with maybe some constraint which produces complementary models. Different decompositions of the strategy space should be also informative of the underlying market dynamics.

We devised a new architecture leveraging the flexibility of DRL and the reliability and efficiency of HC models. We avoided the notoriously hard to explore multi-dimensional action space generally used in the DRL trading literature, by having HC models which do the asset allocation and the DRL agent selects which seems to be the best one (based on recent performance) to act in the real environment. The applications are numerous, with the system being quite flexible and open to further research. One could use various different base models or use different reward functions (or combination thereof) for the DRL. One could even include a human in the loop as an additional low-level agent. Moreover, adding or removing assets does not require DRL retraining since the HC models deal with the actual asset allocation.

Author Contributions: Conceptualization, A.M. and A.E.; methodology, A.M.; software, A.M.; validation, A.M. and A.E.; investigation, A.M.; data curation, A.M.; writing—original draft preparation, A.M.; writing—review and editing, A.E.; supervision, A.E.; project administration, A.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by EPSRC Centre for Doctoral Training in High Performance Embedded and Distributed Systems grant number EP/L016796/1.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in for free as specified in the text.

Acknowledgments: The authors would like to thank Radu Millea and Julian Sutherland for constructive feedback and proofreading the article.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

1 https://huggingface.co/blog/deep-rl-ppo (accessed on 20 June 2022).

2 https://riskfolio-lib.readthedocs.io/en/latest/index.html (accessed on 20 June 2022).

3 https://docs.ray.io/en/latest/rllib/index.html (accessed on 20 June 2022).

4 https://github.com/ray-project/ray/blob/master/rllib/algorithms/ddpg/ddpg.py (accessed on 20 June 2022).

References

Betancourt, Carlos, and Wen-Hui Chen. 2021. Deep reinforcement learning for portfolio management of markets with a dynamic number of assets. *Expert Systems with Applications* 164: 114002. [CrossRef]

Black, Fischer, and Robert Litterman. 1992. Global portfolio optimization. *Financial Analysts Journal* 48: 28–43. [CrossRef]

Bodnar, Taras, Stepan Mazur, and Yarema Okhrin. 2017. Bayesian estimation of the global minimum variance portfolio. *European Journal of Operational Research* 256: 292–307. [CrossRef]

Burggraf, Tobias. 2021. Beyond risk parity—A machine learning-based hierarchical risk parity approach on cryptocurrencies. *Finance Research Letters* 38: 101523. [CrossRef]

Choueifaty, Yves, and Yves Coignard. 2008. Toward maximum diversification. *The Journal of Portfolio Management* 35: 40–51. [CrossRef]

Clarke, Roger, Harindra De Silva, and Steven Thorley. 2011. Minimum-variance portfolio composition. *The Journal of Portfolio Management* 37: 31–45. [CrossRef]

De Prado, Marcos Lopez. 2016. Building diversified portfolios that outperform out of sample. *The Journal of Portfolio Management* 42: 59–69. [CrossRef]

Deng, Yue, Feng Bao, Youyong Kong, Zhiquan Ren, and Qionghai Dai. 2016. Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems* 28: 653–64. [CrossRef]

Goldberg, Lisa R., and Ola Mahmoud. 2017. Drawdown: From practice to theory and back again. *Mathematics and Financial Economics* 11: 275–97. [CrossRef]

Hirsa, Ali, Joerg Osterrieder, Branka Hadji-Misheva, and Jan-Alexander Posth. 2021. Deep reinforcement learning on a multi-asset environment for trading. *arXiv* arXiv:2106.08437.

Jiang, Zhengyao, and Jinjun Liang. 2017. Cryptocurrency portfolio management with deep reinforcement learning. Paper presented at the 2017 Intelligent Systems Conference (IntelliSys), London, UK, September 7–8. pp. 905–13.

Jiang, Zhengyao, Dixing Xu, and Jinjun Liang. 2017. A deep reinforcement learning framework for the financial portfolio management problem. *arXiv* arXiv:1706.10059.

Jurczenko, Emmanuel. 2015. *Risk-Based and Factor Investing*. Amsterdam: Elsevier.

Kolm, Petter N., Reha Tütüncü, and Frank J. Fabozzi. 2014. 60 years of portfolio optimization: Practical challenges and current trends. *European Journal of Operational Research* 234: 356–71. [CrossRef]

Ledoit, Olivier, and Michael Wolf. 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance* 10: 603–21. [CrossRef]

Li, Yang, Wanshan Zheng, and Zibin Zheng. 2019. Deep robust reinforcement learning for practical algorithmic trading. *IEEE Access* 7: 108014–22. [CrossRef]

Li, Yuxi. 2017. Deep reinforcement learning: An overview. *arXiv* arXiv:1701.07274.

Lillicrap, Timothy P., Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv* arXiv:1509.02971.

Maillard, Sébastien, Thierry Roncalli, and Jérôme Teïletche. 2010. The properties of equally weighted risk contribution portfolios. *The Journal of Portfolio Management* 36: 60–70. [CrossRef]

Markowitz, Harry M. 1952. Portfolio Selection. *The Journal of Finance* 7: 77–91.

Michaud, Richard O., and Robert Michaud. 2007. Estimation Error and Portfolio Optimization: A Resampling Solution. Available online: https://ssrn.com/abstract=2658657 (accessed on 20 September 2022).

Millea, Adrian. 2021. Deep reinforcement learning for trading—A critical survey. *Data* 6: 119. [CrossRef]

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, and et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518: 529–33. [CrossRef]

Moody, John, and Matthew Saffell. 1998. Reinforcement learning for trading. In *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, vol. 11.

Mosavi, Amirhosein, Yaser Faghan, Pedram Ghamisi, Puhong Duan, Sina Faizollahzadeh Ardabili, Ely Salwana, and Shahab S. Band. 2020. Comprehensive review of deep reinforcement learning methods and applications in economics. *Mathematics* 8: 1640. [CrossRef]

Raffinot, Thomas. 2017. Hierarchical clustering-based asset allocation. *The Journal of Portfolio Management* 44: 89–99. [CrossRef]

Raffinot, Thomas. 2018. The hierarchical equal risk contribution portfolio. Available online: https://ssrn.com/abstract=3237540 (accessed on 20 September 2022).

Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv* arXiv:1707.06347.

Schulman, John, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. Paper presented at the International Conference on Machine Learning, Lille, France, July 6–11. pp. 1889–97.

Théate, Thibaut, and Damien Ernst. 2021. An application of deep reinforcement learning to algorithmic trading. *Expert Systems with Applications 173*: 114632.

Uryasev, Stanislav. 2000. Conditional value-at-risk: Optimization algorithms and applications. Paper presented at the IEEE/IAFE/INFORMS 2000 Conference on Computational Intelligence for Financial Engineering (CIFEr) (Cat. No. 00TH8520), New York, NY, USA, March 28, pp. 49–57.

Wei, Haoran, Yuanbo Wang, Lidia Mangu, and Keith Decker. 2019. Model-based reinforcement learning for predictions and control for limit order books. *arXiv* arXiv:1910.03743.

Yang, Hongyang, Xiao-Yang Liu, Shan Zhong, and Anwar Walid. 2020. Deep reinforcement learning for automated stock trading: An ensemble strategy. Paper presented at the First ACM International Conference on AI in Finance, New York, NY, USA, October 15–16. pp. 1–8.

Yu, Pengqian, Joon Sern Lee, Ilya Kulyatin, Zekun Shi, and Sakyasingha Dasgupta. 2019. Model-based deep reinforcement learning for dynamic portfolio optimization. *arXiv* arXiv:1901.08740.

Zejnullahu, Frensi, Maurice Moser, and Joerg Osterrieder. 2022. Applications of reinforcement learning in finance–trading with a double deep q-network. *arXiv* arXiv:2206.14267.

Zhang, Zihao, Stefan Zohren, and Stephen Roberts. 2020. Deep reinforcement learning for trading. *The Journal of Financial Data Science* 2: 25–40. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

International Journal of
Financial Studies

MDPI

Article

GALSTM-FDP: A Time-Series Modeling Approach Using Hybrid GA and LSTM for Financial Distress Prediction

Amal Al Ali [1,†], Ahmed M. Khedr [2,3,*,†], Magdi El Bannany [4,5,†] and Sakeena Kanakkayil [2,†]

[1] Information Systems Department, University of Sharjah, Sharjah 27272, United Arab Emirates
[2] Computer Science Department, University of Sharjah, Sharjah 27272, United Arab Emirates
[3] Mathematics Department, Zagazig University, Zagazig 44519, Egypt
[4] College of Business Administration, Umm Al Quwain University, Umm Al Quwain 536, United Arab Emirates
[5] Department of Accounting and Auditing, Faculty of Business, Ain Shams University, Cairo 11566, Egypt
* Correspondence: akhedr@sharjah.ac.ae
† These authors contributed equally to this work.

Abstract: Despite the obvious benefits and growing popularity of Machine Learning (ML) technology, there are still concerns regarding its ability to provide Financial Distress Prediction (FDP). An accurate FDP model is required to avoid financial risk at the lowest possible cost. However, in the Internet era, financial data are exploding, and they are being coupled with other kinds of risk data, making an FDP model challenging to operate. As a result, researchers presented several novel FDP models based on ML and Deep Learning. Time series data is are important to reflect the multi-source and heterogeneous aspects of financial data. This paper gives insight into building a time-series model and forecasting distress far in advance of its occurrence. To build an efficient FDP model, we provide a hybrid model (GALSTM-FDP) that incorporates LSTM and GA. Unlike other previous studies, which established models that predicted distress probability only within one year, our approach predicts distress two years ahead. This research integrates GA with LSTM to find the optimum hyperparameter configuration for LSTM. Using GA, we focus on optimizing architectural aspects for modeling the optimal network based on prediction accuracy. The results showed that our algorithm outperforms other state-of-the-art methods in terms of predictive accuracy.

Keywords: financial distress prediction (FDP); long short term memory (LSTM); genetic algorithm (GA); machine learning (ML)

1. Introduction

Financial distress prediction (FDP), the significant factor of enterprise risk management, is also the core of enterprise financial distress theory (Wanke et al. 2015). As a result of the COVID-19 pandemic and current global economic recession, the probabilities of numerous sorts of businesses entering financial distress or insolvency is steadily increasing (El-Bannany et al. 2021; Khan et al. 2020). Many studies have shown that ignoring enterprise financial risks is an important cause of business failure, and FDP has a great impact on corporate sustainability (Mehreen et al. 2020). Thus, for enterprises with significant financial risks, but which have not yet caused significant losses, it is critical to be able to timely spot potential financial difficulties and warn them (Hu and Sathye 2015; Khedr et al. 2021). For businesses and other market participants, obtaining an FDP model with significant operability, strong predictive accuracy, and a broad application scope is crucial. Machine Learning (ML) approaches began to emerge as efforts in the field of FDP increased, including Neural Network (NN), Genetic Algorithm (GA), Support Vector Machines (SVM), etc. Moreover, as ML technology improves, combining several ML approaches during the FDP process has become a new trend (Sreedharan et al. 2020a; Zhu et al. 2022). The majority of studies indicate that data mining techniques can forecast distress better and outperforms

traditional methods (Sreedharan et al. 2020b). For model construction, on one hand, classic statistical and ML methods are applied in feature engineering and classification, such as Naïve Bayesian, SVM, and ensemble learning, including decision-trees-based Gradient Boosting Decision Tree (GBDT), Random Forest (RF), eXtreme Gradient Boosting (XGB), and Adaptive Boosting (AdaBoost) (Başoğlu Kabran and Ünlü 2021; Gepp and Kumar 2015; Khedr et al. 2021; Kim et al. 2019; Min and Lee 2005; Wu et al. 2008). On the other hand, various Deep Learning (DL) models are also employed for modeling, such as Genetic Algorithm (GA), Convolutional Neural Network, and Self Organizing Map (SOM) (Bukhari et al. 2020; Jang et al. 2019). Because of their excellent prediction accuracy, Neural Networks are frequently employed in FDP (Zhu et al. 2022).

Past studies on FDP show that DL techniques performed better than traditional techniques over time-series data (El Bannany et al. 2021; Geng et al. 2015; Ribeiro and Lopes 2011). In particular, among DL architectures, Recurrent Neural Networks (RNNs) have proved their superiority in analyzing financial time-series data (Bukhari et al. 2020). A recurrent neural network (RNN) is an ANN in which node connections form a directed graph along a time sequence. As a result, it can display temporal dynamic behavior. The selection of hyperparameters has a significant impact on the efficiency of an LSTM. The fine-tuning of these parameters is a critical step in increasing the model's prediction accuracy. Using brute force trial and error to try every possible combination of reasonable parameters is one method for determining ideal hyperparameters. Many LSTMs must be trained one by one, multiple times. This takes a significant amount of time and computer resources.

This paper gives insight into building a time-series model and forecasting distress far in advance of its occurrence. In order to make up for the shortcomings of a single prediction model, a new hybrid model that incorporates LSTM and GA is presented in this work. We created an FDP model using a DL algorithm called the Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN). An LSTM network is a type of deep RNN model composed of LSTM units, which was introduced as a way to overcome the long-term dependency problem (Hochreiter and Schmidhuber 1997). The LSTM can scale to much longer sequences than simple RNN, overcoming the intrinsic drawbacks of simple RNN. The LSTM is advantageous for forecasting financial distress because it can efficiently learn sequential patterns in the given data including sequential or temporal characteristics and predicts time-series data well. GA is a metaheuristic and stochastic optimization algorithm inspired by the process of natural evolution, and it is widely used to find near-optimal solutions to optimization problems with large search spaces (Sun and Hui 2006). To build an efficient FDP model, we provide a hybrid model that incorporates LSTM and GA. In contrast to other previous studies that predicted financial distress probability only within one year, our GALSTM-FDP approach predicts distress two years ahead. This research integrates GA with LSTM to find the optimum hyperparameter configuration for LSTM. Using GA, we focus on optimizing architectural aspects for modeling the optimal network based on prediction accuracy. GA is a heuristic search and optimization technique based on natural selection, and it is commonly employed to obtain a near-optimal solution to optimization problems with a broad parameter space (Sun and Hui 2006). To the best of the authors' knowledge, this is the first study that uses a hybrid optimized GA and LSTM model in FDP studies based on time-series data and predicts distress more than a year in advance. Because we incorporated just within-company data, our model is compatible with any region data.

The major contributions of this paper are: First, we propose an LSTM-RNN model as an FDP model that accurately predicts distress two years ahead. The LSTM is one of the most advanced DL architectures for capturing long-term dependencies from financial time-series data. In order to improve the prediction performance of the FDP model, we provide a hybrid optimized GALSTM-FDP model. We integrate GA with LSTM to find the optimum hyperparameter configuration for LSTM where we focus on optimizing architectural aspects of the model for enhanced prediction power. Second, we provide a

strong and effective distress prediction model that is beneficial for the corporate, market participants and policymakers.

The rest of the study is structured as follows: Section 2 provides a brief overview of the related research. The proposed GALSTM-FDP model formation is presented in Section 3. Data and modeling details are provided in Section 4. Section 5 discusses the results and compares the prediction performance of the proposed hybrid model with existing ML models. Finally, Section 6 concludes the paper by stating potential future directions.

2. Related Research

Statistical and ML approaches have been utilized in earlier research to forecast financial distress (Sreedharan et al. 2020a). Conventional statistical models employed in the domain of distress prediction include discriminant analysis and the logit model. Altman (1968) utilized discriminant analysis to anticipate financial distress in firms, whereas Ohlson (1980) used the logit model to forecast financial distress. Later, Falbo (1991) presented a modified discriminatory model of study over a period of several years using financial stability ratios to enhance discriminatory strength. Traditional linear techniques were ineffective and had their own limits. Such straightforward approaches need more predictive history data and could not be utilized to develop a strong classifier for actual predictions. SVM as a non-parametric binary classification technique is widely used in financial time-series forecasting, and the work in (Başoğlu Kabran and Ünlü 2021) utilized SVM for the prediction of bubbles.

As financial data are considered non-linear, statistical approaches cannot be utilized to build a credible prediction model. ML methods are widely employed in financial distress prediction because of their significant benefits in extracting non-linear data relationships without previous input knowledge. As a result, ML techniques such as Logistic Regression, SVM (Min and Lee 2005; Sun et al. 2017), and NNs (Cleofas-Sánchez et al. 2016; Ravisankar and Ravi 2010) were incorporated in past research. According to their results, these approaches have proven relative effectiveness over statistical methods due to their capability to identify noisy data without making any statistical predictions. Cleofas-Sanchez et al. applied Santiago-Montero's hybrid associative classifier with translation (HACT) model to predict financial distress and provided empirical results supporting that HACT dominated four traditional neural networks, including multi-layer perceptron (MLP), radial basis function (RBF), Bayesian network (BN), and voted perceptron (VP), one SVM, and one multi-variate logistic regression (LR) model (Cleofas-Sánchez et al. 2016). While the HACT model can be trained easily due to its feed-forward learning framework, it is only suitable for simple data with repetitive structures. It would not be able to generate good learning and prediction results for complicated data. Chou et al. proposed a GA-based fuzzy clustering algorithm for FDP (Chou et al. 2017). In particular, key financial ratios selected by the GA are clustered by the fuzzy C-means clustering after the training data are divided into financially distressed and financially non-distressed samples. The optimal number of clusters for both samples are decided by the WB index. However, the major drawbacks of this method are the high possibility of over-fitting the training data as well as the time consumed to find the optimal results. Ruibin et al. evaluated the effectiveness of ML approaches for predicting distress in publicly listed Chinese businesses (Geng et al. 2015). They examined three popular classifiers in data mining and analyzed their effectiveness using majority voting. A study on datasets gathered from various nations is also observed utilizing data mining techniques in (Bae 2012). SVM and NN classifiers show fairly good efficiency in predicting financial distress.

The learning time of NN classifiers was substantially decreased with the introduction of parallel processing technology. As a result, researchers have attempted to form DNN-based models for prediction purposes (Huang and Yen 2019; Shen et al. 2015). Due to their exceptional classification performance, DL algorithms were eventually used in the field of FDP (Glorot et al. 2011). Ribeiro and Lopes presented a DBN-based prediction model to forecast failures in French companies (Ribeiro and Lopes 2011). Matin et al. (2019) present distress prediction using DL, which employs unstructured textual information

in financial statements for prediction. To improve prediction performance on corporate financial problems, a genetic algorithm was combined with existing machine learning techniques. Jie Sun used the genetic algorithm in conjunction with the Decision Tree to optimize the financial ratio defined in the prediction of distress (Sun and Hui 2006). Hou, in 2016, introduced the K-clustering algorithm based on the genetic algorithm to solve certain conventional K-means clustering problems whilst predicting financial distress (Hou 2016). Kim et al. developed a genetic approach to optimize several heterogeneous design variables of SVMs at the same time (Kim et al. 2019). Though deep NN models are a great tool for predicting financial difficulties, there are numerous disadvantages to using them. The capacity of NN models to describe the final decision that models obtain is limited. Although NNs can give a substantial solution to the goal problem, they cannot provide a good explanation for the projected results. Shin and Lee presented a hybrid strategy of merging ANN and evolutionary algorithms and extracting rules from the bankruptcy prediction model to overcome this difficulty (Shin and Lee 2002). Another issue is that, like other NN models, an RNN contains numerous parameters that the researcher must tune. However, due to time and computational constraints, it is not possible to skim over a parameter space in order to find the optimal setting. Jang et al. (2019) developed a business failure prediction model based on LSTM and showed that it outperforms FNN and SVM in predicting construction contractors' business failure. Halim et al. (2021) examined the effectiveness of DL models such as RNN, GRU, and LSTM for FDP among publicly listed organizations in Malaysia using time-series data only in a single year. El-Bannany et al. (2020) employs three distinct DL models: MLP, LSTM, and CNN for FDP considering single-year data.

From the previous studies, it is evident that not much research has been carried out considering the problem as a time series to capture long-term dependencies from financial time-series data. We have selected an effective algorithm for time-series analysis, the LSTM-RNN algorithm. In order to make up for the shortcomings of a single prediction model, a new hybrid model that incorporates LSTM and GA is presented in this work. To the best of authors' knowledge, this is the first study that uses a hybrid optimized GA and LSTM model in FDP studies based on time-series data and predicts distress more than a year in advance. Because we only incorporated within-company data, our model is compatible with any region data. In contrast to other previous studies that predicted financial distress probability only within one year, this paper gives insight into building a time-series model and forecasting distress far in advance of its occurrence. Using GA, we focus on optimizing architectural aspects for improved performance. By adjusting the hyperparameters, we use the evolutionary algorithm approach to model the optimal network based on prediction accuracy.

3. Proposed GALSTM-FDP Model Formation

In this section, we provide an overview of the main models used in this study, including the LSTM, GA, and the proposed hybrid GALSTM-FDP model. The goal of this study is to identify the optimal LSTM parameter values for predicting financial distress using a genetic algorithm. We concentrate on fine-tuning the hyperparameters—the number of hidden layers, neurons, epochs, and batch size. Following this phase, the top five networks are identified and assessed based on prediction accuracy. The suggested hybrid model is then compared to traditional machine learning techniques in the following part of the research.

3.1. Long Short-Term Memory (LSTM)

Hochreiter and Schmidhuber proposed the LSTM neural network, which is commonly used to process sequence information due to its benefits in recognizing long-term dependencies (Hochreiter and Schmidhuber 1997). As a result, creating an LSTM model for financial time-series data is theoretically possible. An LSTM network is a type of deep RNN model composed of LSTM units. As discussed earlier, RNN is a DL network with internal feedback between neurons. The structure of RNN, as shown in Figure 1, can

theoretically map from all prior inputs to each output. The recurrent connections in RNN allows the memory of earlier inputs to survive in the internal state of the network and therefore impact the output of the network. Unfortunately, the RNN struggles from the vanishing gradients issue, that is, the impact of an input datum on a hidden layer and thus on the network output, which either deteriorates or explodes exponentially while looping through the recurrent connections of the network. Hochreiter and Schmidhuber (1997) advocated for LSTM for a lengthy span of time to address this issue. Its architecture is made up of LSTM memory blocks that are comparable to the RNN's hidden neurons. They are capable of learning long-term dynamics and avoiding the disappearing and inflating gradient issues. The LSTM-RNN can understand temporal and sequential patterns from time-series or data sequences. Many recent research have used LSTM to forecast time-series data and demonstrated that it outperforms other methodologies. To anticipate the sudden stochastic fluctuation of the financial market, a novel hybrid approach with the power of fractional order derivative was provided with the properties of LSTM networks in (Bukhari et al. 2020).

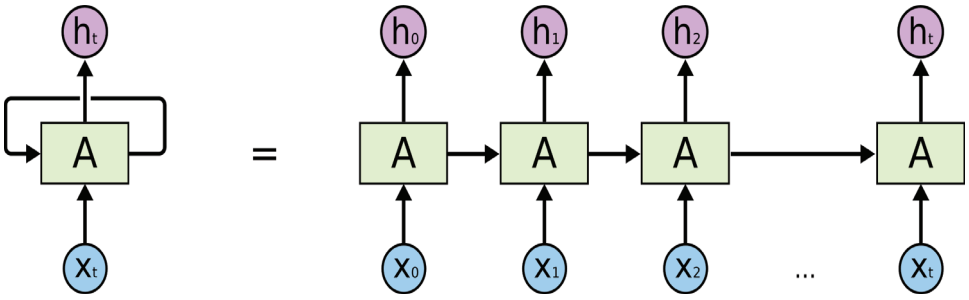


Figure 1. Structure of an RNN.

3.2. Genetic Algorithm (GA)

GA is a metaheuristic and stochastic optimization algorithm inspired by the process of natural evolution (Sun and Hui 2006). They are widely used to find near-optimal solutions to optimization problems with large search spaces. GA is a powerful optimization technique that works on principles such as selection, crossover, and mutation. Processing the GA can be divided into different stages: initialization, fitness calculation, termination condition check, selection, crossover, and mutation.

3.3. Hybrid GALSTM-FDP Model for FDP

This study develops an FDP model using an LSTM-RNN coupled with GA to obtain the best parameter setting for our problem. As previously stated, an LSTM is a form of RNN designed to learn sequential and temporal patterns from time-series data. The proposed GALSTM-FDP structure contains one input layer with time step = 3, one or more hidden layers, a dropout layer for each LSTM layer, and an output layer. The flow diagram of the proposed model is given in Figure 2.

A dropout layer should be added to every LSTM layer. By neglecting randomly chosen neurons during training, this layer alleviates over-fitting and therefore decreases sensitivity to the particular weights of individual neurons. The 20% rule is frequently employed as a reasonable balance between maintaining model accuracy and avoiding over-fitting.

We have used a tumbling window rather than a sliding window as our data are panel data that consider each company distinctly. For this, we have set our stride parameter to 5 and window length to 3. As we are predicting distress two years ahead, the number of input is 3 (taking the first 3 years in the dataset and skipping the next year) with 26 features.

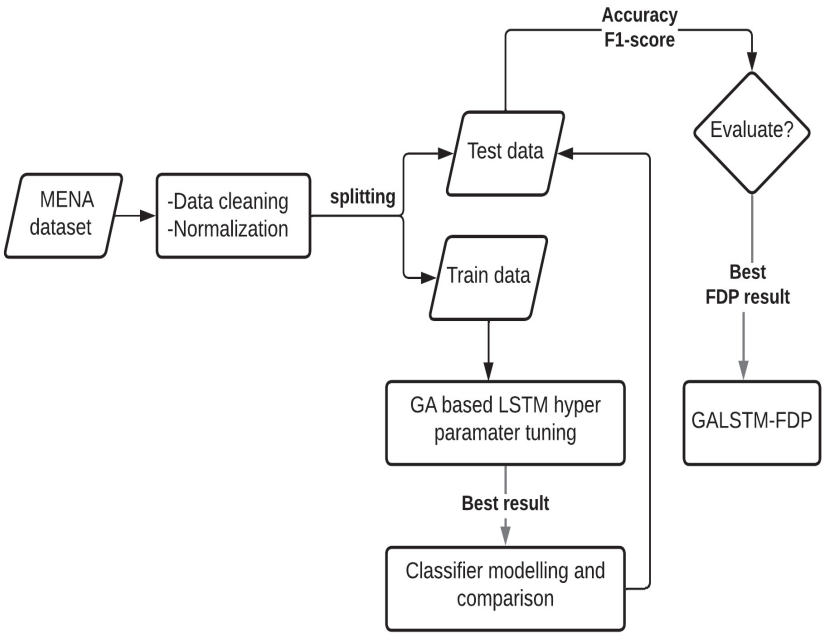


Figure 2. Flow diagram of the GALSTM-FDP model for FDP.

Optimizing the LSTM Model Using GA

This study examines how GAs can be used to optimize LSTM networks for FDP. We have applied GA to optimize the number of units in the LSTM layers, activation functions, number of epochs, and batch size. The algorithm starts with a random initial population, forming the core of the GA. At each generation, a new population is created from the initial population after the process of evaluation, selection, crossover, and mutation. This process continues over the generations until an optimum solution is attained. Figure 3 depicts the process of optimizing the LSTM model using GA.

- Generate the initial population: The creation of the initial population, which forms the basis of GA is the first step. By randomly mixing various hyperparameter values, 20 distinct LSTM networks are produced, constituting the starting population of the GA. Antonio Dourado (2013) has offered a solid rule of thumb for avoiding overfitting: start with [Number of Training Samples/2 $\times$ (Number of Input Neurons + Number of Output Neurons). In most cases, it works, but if the problem is too simple or complex, we can experiment increasing or decreasing the number of neurons. We have picked two values that are less than 64 and two values that are larger than 64. The popular Adam optimizer was chosen as the activation function for this model. The initial population in this study is made up of a collection of 20 random networks. The other two parameters that can be varied are the number of epochs and batch size.
- Compute fitness: Each individual network in the population is trained and tested using the MENA dataset and scored according to the predictive accuracy. We did not run cross-validation on this dataset because it is a time-specific dataset. The highest-scoring networks are retained in order to increase the population of the next generation. The remaining networks of the existing population are discarded.
- Selection: An LSTM network with high accuracy has a greater probability of being chosen for the following generation. During the selection phase, we identify networks from the present population that will be passed down unchanged to the future generation. The top-ranked quarter (five numbers) networks based on accuracy scores are transferred directly to the next population. In order to prevent being trapped at the local maximum, three low-performing networks are also preserved and transferred to the next generation.

- Breeding: For the next generation, we currently have eight networks in the population. The remaining 12 are the product of crossover or breeding. To produce one or more children for the following generation, two networks from the current population, known as parents, are required. The parents are chosen based on their scores, and the network parameters are combined to generate a new offspring that is a hybrid of its parents. Each child in this study is a network with a random set of parameters from its parents.
- Mutation: We randomly adjust some of the properties of random networks in the population to have a population for the next generation. This technique aims to create better networks for the population.

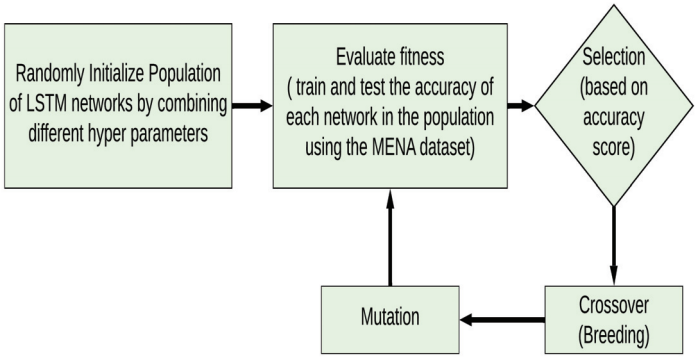


Figure 3. Optimizing the LSTM model using GA for FDP.

Algorithm 1 involved in this study is given below:

Algorithm 1 GALSTM-FDP

1: *Construct a population of LSTM networks and allocate them all random hyperparameters.*
   - Number of hidden layers: {1, 2, 3, 4}
   - Neurons per layer: {16, 32, 64, 128, 256}
   - Batch size: {35, 70, 105, 140, 175 }
   - Epochs: {50, 100, 200, 300, 400, 500}
   - Select 20 random LSTM as the current population.
2: *Train each network in the population using the given MENA dataset.*
3: *Calculate the prediction accuracy for each one and pick the top performing ones for the next step.*
4: *Select some top networks and non-top networks (to avoid getting stuck at local maximum).*
5: *Crossover: The parameters of two members of the chosen networks are crossed over. By merging the parents, this will result in a "child" network possessing some features of the first and some of the second.*
6: *Mutate the parameters of some of the child networks*
7: *Save the generated child networks in a new population and allocate it to the variable that holds the former population.*
8: *Repeat steps 2 to 7 for 10 generations.*

We used an initial set of 20 random networks to establish the population in this study, and we repeated the evaluation, selection, crossover, and mutation procedure for 10 generations. As a result of our research, we have trained approximately 200 LSTMs, resulting in a stronger population over time. In the following part, the top five LSTM networks from the final population set are determined and examined based on prediction accuracy. The predictive accuracy of networks in the population improved over generations, and the models were trained and generated using the Python libraries Scikit-learn and Keras. We also evaluated the performance of the proposed model to that of traditional ML models. The simulation results revealed that the proposed hybrid model has much greater prediction accuracy when compared to SVM, DT, and standalone LSTM classifier algorithms.

4. Data and Modeling

The goal of this study is to identify the optimal LSTM parameter values for predicting financial distress using a genetic algorithm. We have applied GA to optimize the number of units in the LSTM layers, activation functions, number of epochs, and batch size. Following this phase, the top five networks are identified and assessed based on prediction accuracy. The suggested hybrid model is then compared to traditional machine learning techniques in the following part of the research. To evaluate whether a company shows up as financially distressed or not, the FDP problem discussed in this paper is modeled as a binary classification problem. In the financial dataset, the output or target attribute belongs to any one of the two classes: financially distressed companies and financially healthy companies. All of the other attributes in the dataset are continuous values with the exception of this target attribute, which has a binary value.

Dataset Description: The dataset considered for evaluating the predictive performance includes sample data gathered from MENA (Middle East and North Africa) listed firms, extracted from the Osiris database. Financial companies are exempted due to the difference in their operational environment. We selected only the companies with 5 years (2015–2019) of consecutive data in the final dataset, totaling to 9765 company years of 1953 companies. The financial variables extracted from financial statements and balance sheets of respective companies are selected considering the variables and ratios used in prior studies in this area. A detailed description of the ratios calculated from 20 financial variables is given in Table 1. The 20 variables along with 6 ratios are taken as input to the network.

Table 1. Financial Indicators.

Indicators	Formula
Solvency	Total liabilities/Total assets Current assets/Current liabilities Current assets-inventory/Current liabilities Total liabilities/total shareholders' equity Current liabilities/total assets Net operating cash flow/current liabilities Earnings before interest and tax/interest expense
Capital Expansion	Net profit/number of ordinary shares at the end of year Net assets/number of ordinary shares at the end of year Net increase in cash and cash equivalents/number of ordinary shares at the end of year Capital reserves/number of ordinary shares at the end of year
Profitability	(Sales revenue–sales cost)/sales revenue Net profit/sales revenue Earnings before income tax/average total assets Net profit/average total assets Net profit/average current assets Net profit/average fixed assets Net profit/average shareholders' equity
Business Development	Business income of this year/Business income of last year Total assets of this year/total assets of last year Net profit of this year/net profit of last year
Operational Capabilities	Main business income/average total assets Sales revenue/average current assets Sales revenue/average fixed assets Main business cost/average inventory Main business income/average balance of accounts receivable Cost of sales/average payable accounts

Table 1. *Cont.*

Indicators	Formula
Structural Soundness	Current assets total assets Fixed assets/total assets Shareholders' equity/fixed assets Current liabilities/total liabilities

Prior to actual fitting of the model to the dataset, three steps must be completed: (1) Data cleaning, (2) Normalization, and (3) Splitting.

1. Data cleaning: This phase is used to validate the data contained within the dataset. As seen in Table 2, the value ranges between variables have a wide variation. Our samples, like most real datasets, contain null values and missing properties. These issues were resolved during the pre-processing stage, where we scaled the dataset after filling in the missing values with forward and backward fill within each company as the units of several variables were also different.
2. Normalization: The model rescales the variables to a range of −1 to 1 to produce a fair result. This process is needed to convert all of the column values in the dataset to a common scale without distorting the value ranges or causing data loss.
3. Dataset samples splitting: The dataset was then turned into the time-series data that the LSTM network required. For a realistic evaluation of performance, datasets must be divided into testing and training sets. We cannot partition our data like other non-time specific datasets because they are time series. At any given time, our network receives 5 years' worth of data (first 3 years of data from 5 years as input and fifth year as output). Considering this, we have divided our training set into a multiple of 5: 7810 samples in the training set and the remaining 1955 years in the testing set. This is approximately 80% of data in the training samples and the remaining 20% in the testing samples.

Table 2. Descriptive statistics of all individual financial attributes.

Variables	Mean	Min	Max
Total Liabilities	88,751.36508	13	1,669,220
Total Assets	183,280.635	5763	3,091,702
Current Assets	43,249.87	138	676,520
Current Liabilities	46,492.642	10	790,074
Accounts Receivable	11,041.079	5.63	200,071
Accounts Payable	11,966.0634	205	374,494
Total Shareholders Equity	93,131.111	5565	1,422,482
Net Cash Flow	17,297.667	−4600	481,539
EBIT	8274	16	267,461
Net Profit	4253.9524	−31,571	156,702
Cash and Cash Equivalent	12,907.0158	19	209,716
Cost of Goods Sold	22,169.1746	345	266,764
Sales	1,730,068.53	93	10,301,478
Shares	402,708	30,318.8	3,901,347
Capital Reserves	5774.65	−615	182,827
Fixed Assets	140,030.76	599	2,415,182
Average Total Assets	274,315.7778	8815	4,608,537.5
Average Current Assets	66,022.484	210	1,048,037
Average Fixed Assets	208,293.2937	825	3,560,500
Average Equity	140,519.53	8476	2,227,284
Average Accounts Receivable	13,773.69	4069	327,807
Average Accounts Payable	16,861.05	794	555,164.5
Net Increase in Cash	−3622.984	−151,247	16,129

5. Empirical Analysis

Our model is built on the popular Keras ML library in Python. We have used a tumbling window rather than a sliding window as our data are panel data that consider each company distinctly. For this, we have set our stride parameter to 5 and window length to 3. As we are predicting distress two years ahead, the number of input is 3 (taking the first 3 years in the dataset and skipping the next year) with 26 features.

To evaluate whether a company shows up as financially distressed or not, the FDP problem discussed in this paper is modeled as a binary classification problem. In the financial dataset, the output or target attribute belongs to any one of the two classes: financially distressed companies and financially healthy companies. All of the other attributes in the dataset are continuous values with the exception of this target attribute, which has a binary value.

In this research, *negative* are companies that are financially distressed, whereas *positive* are those that are financially sound. If the F1-score is employed as a performance metric, a balance between precision and recall could be attained.

5.1. FDP Performance Evaluation Metrics

We compare the classifiers' performances in predicting financial distress with two common ML evaluation metrics:

1. Accuracy: For performance assessment, the accuracy of the test data is considered. Accuracy is the percentage of the number of correct predictions from all the predictions made.
2. F1-score: F1-score is a function of precision (Equation (1)) and recall (Equation (2)) as defined by Equation (3).

$$Precision = \frac{True - Positive}{(True - Positive + False - Positive)} = \frac{True - Positive}{(Total - Predicted - Positive)} \quad (1)$$

$$Recall = \frac{True - Positive}{(True - Positive + False - Negative)} = \frac{True - Positive}{(Total - Actual - Positive)} \quad (2)$$

$$F1 - score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (3)$$

5.2. Correlation and Feature Importance

The descriptive statistics of the attributes are given in Table 2. Figure 4 shows the Pearson's correlation, and Figure 5 depicts the feature importance of the attributes, respectively. A high correlation means values between -0.50 and -1.00. Feature importance order is depicted in Figure 5. We included all the attributes, as omitting the least important ones resulted in lower performance; extreme outliers in the dataset were detected, analyzed (e.g., to determine whether they resulted from errors in data entry), and removed.

5.3. Analysis of the Proposed Hybrid GALSTM-FDP Model for FDP

This section presents and analyzes the findings of the proposed hybrid GALSTM-FDP model for FDP. The average accuracy values of all 20 models in the population over 10 generations are shown in Figure 6.

According to the graph, the financial distress prediction accuracy of networks in the population has increased over generations. This means that the GA optimizes the network at each iteration, resulting in the best solution for FDP at the completion of the iterations. At each generation, a set of 20 networks is trained and scored, then crossed over and mutated to produce the new population of networks for the coming generation. Table 3 displays the parameters of the five highest scoring networks based on the percentage of prediction accuracy and F1-score. Figure 7 depicts the accuracy and F1-score graphs for the same.

The highest predictive accuracy and F1-score for the dataset in percentages are 91.50 and 92.67, respectively. A dataset with a greater number of attributes can be trained more effectively and provide a more robust model than a dataset with fewer attributes. From the

above table, we can see that although we have included three and four hidden layers for analysis the best performing was two layers. That is, an LSTM with two hidden layers is enough for our FDP problem. Training the model for more number of epochs did yield a higher accuracy. Our input layer had 26 input attributes, which in turn is the shape of the input layer. Moreover, 64 or 128 neurons per layer gave a better performance than the others. We set the batch sizes as multiples of 5 because we take 5 years of data as a single input; 35 or 70 inputs per batch performed well in our setting.

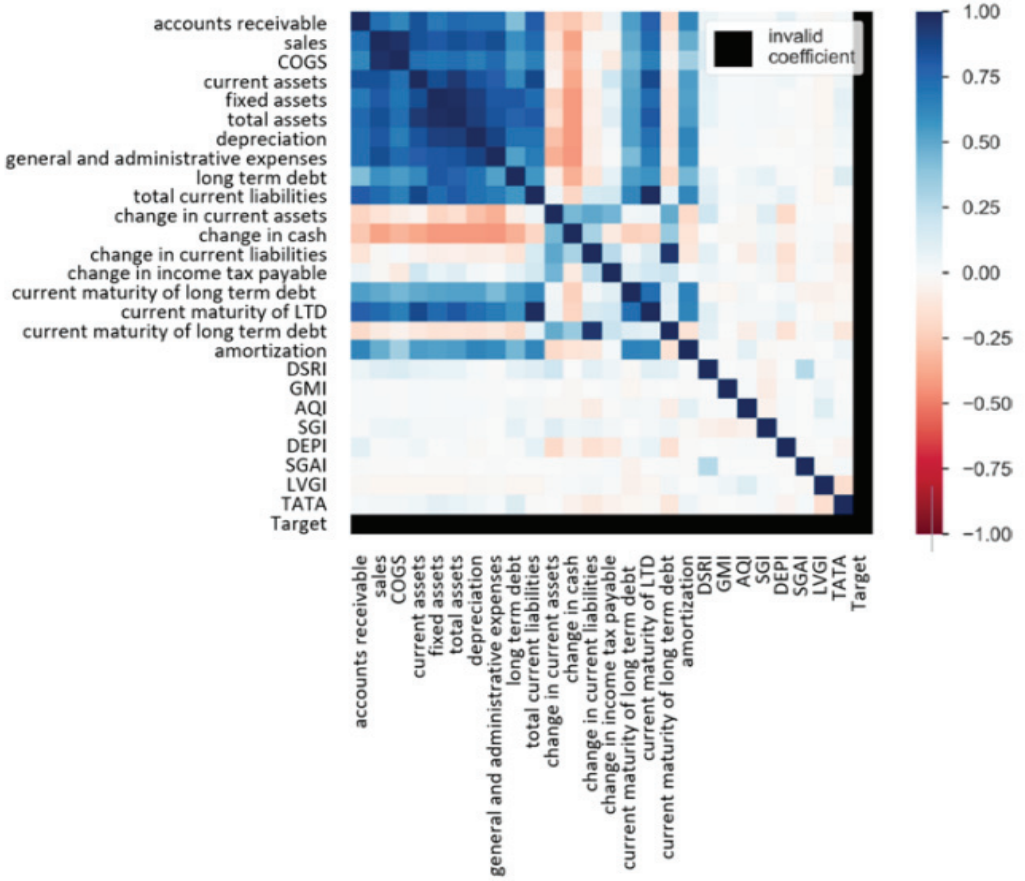


Figure 4. Pearson's r correlation.

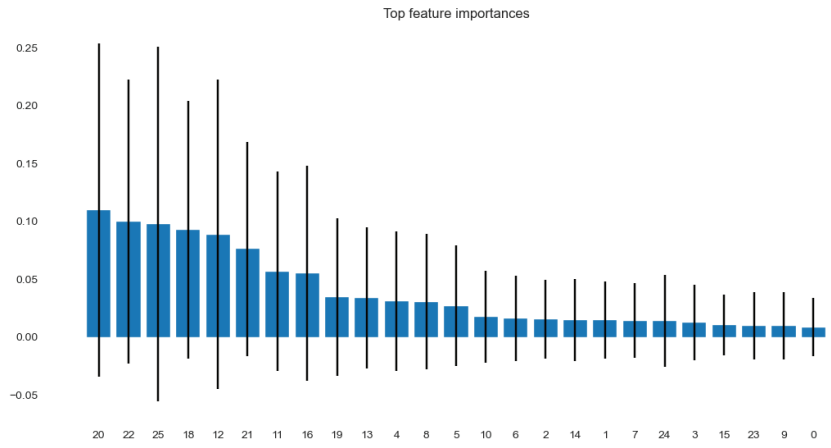


Figure 5. Feature importance of the attributes.

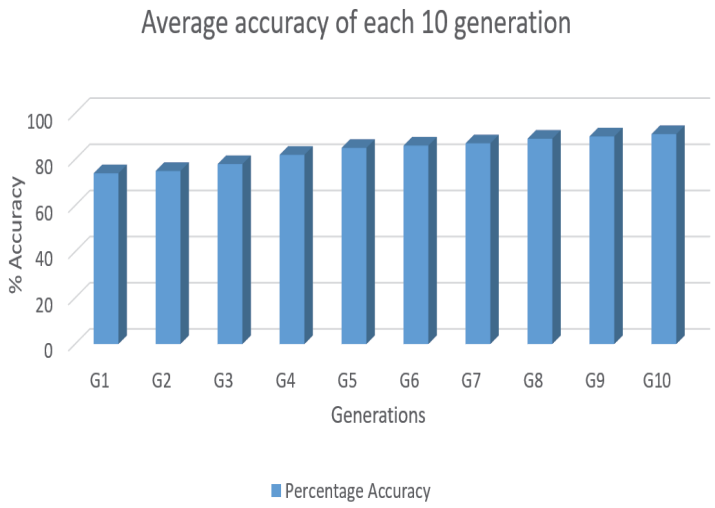


Figure 6. Average accuracy score of the GALSTM-FDP model across 10 generations.

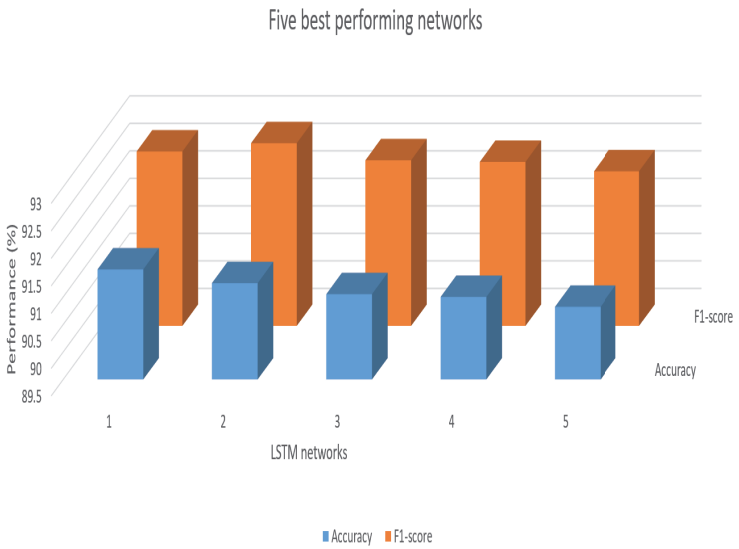


Figure 7. Best-performing 5 LSTM models.

Table 3. GALSTM-FDP: Network parameters of top 5 networks based on percentage of accuracy and F1-score.

Network	Hidden Layers	Neurons per Layer	Batch Size	Epoch	Testing Accuracy (%)	Testing F1-Score (%)
1	2	64	35	500	91.50	92.67
2	2	64	70	400	91.25	92.82
3	1	64	35	500	91.05	92.51
4	2	128	35	400	91.00	92.48
5	1	128	70	500	90.82	92.31

In the last stage of the research, we compared our improved LSTM network against SVM, DT, and basic LSTM classifier algorithms. Prediction outcomes in terms of accuracy and F1-score are shown in Table 4. The statistical findings show that the suggested optimized GALSTM-FDP model's prediction accuracy is substantially greater than that of the traditional ML models.

Table 4. Comparison of GALSTM-FDP, LSTM, SVM, and DT in solving the FDP problem.

	Financial Distress Prediction Performance	
Classifiers	**Accuracy**	**F1-Score**
GALSTM-FDP	91.50	92.67
LSTM	88.75	90.46
SVM	86.50	89.67
DT	81.75	85.46

5.4. GALSTM-FDP: Overall Analysis

The GALSTM-FDP model's greater performance can be attributed to the fact that the architecture of the LSTM network substantially improved learning efficiency and reduced the need for unnecessary computations. We integrated GA with LSTM to find the optimum hyperparameter configuration for LSTM with the aim of optimizing architectural aspects of the model for enhanced prediction power. The application of GA to investigate the optimal architectural factors derives results through this genetic search. The findings imply that proper parameter tuning is a necessary precondition for achieving adequate performance. Even if deep learning algorithms are expanding quickly, finding the ideal set of deep architecture parameters requires extensive understanding. The experimental findings, however, indicate the potential for the hybrid GALSTM-FDP model's application in FDP and show that it can be a useful tool for finding the best or nearly best results. Because we incorporated just within-company data, our model is compatible with any region data. In other words, it is most effective in handling FDP based on time-series data, can accurately predicts distress two years ahead, and is beneficial for the corporate, market participants and policymakers.

6. Conclusions and Future Work

Finding the best-performing FDP model has always been a focus of researchers, and numerous FDP models have been developed since then. We presented a hybrid GALSTM-FDP model for FDP and assessed the performance of LSTM by adjusting the hyperparameters using a GA. It was observed that if the model is configured with one or two hidden layers, an acceptable predicted performance rate can be obtained. We trained and evaluated the models with a dataset of 26 input variables collected from 1953 firms in the MENA area. The simulation results show that the suggested model outperforms standalone LSTM and other conventional ML methods such as SVM and DT. All of the variables in our analysis are within-firm variables. To improve the model, macroeconomic and other industrial aspects may be incorporated in future study.

Author Contributions: All authors contributed together to realize this work, and their exact contributions are difficult to specify. Conceptualization and methodology by A.A.A., S.K., and A.M.K.; Software and original draft preparation by A.A.A. and M.E.B.; Validation, writing, reviewing, editing, and supervision by M.E.B., S.K., and A.M.K. All authors have read and agreed to the published version of the manuscript.

Funding: No funding was obtained for this study.

Informed Consent Statement: Not applicable.

Data Availability Statement: Available on request.

Conflicts of Interest: The authors declare that they have no competing interests.

References

Altman, Edward I. Altman. 1968. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance* 23: 589–609. [CrossRef]

Bae, Jae Kwon Bae. 2012. Predicting financial distress of the South Korean manufacturing industries. *Expert Systems with Applications* 39: 9159–65. [CrossRef]

Başoğlu Kabran, F., and Kamil Demirberk Ünlü. 2021. A two-step machine learning approach to predict S&P 500 bubbles. *Journal of Applied Statistics* 48: 2776–94.

Bukhari, Ayaz H., Muhammad Asif Zahoor Raja, Muhammad Sulaiman, Saeed Islam, Muhammad Shoaib, and Poom Kumam. 2020. Fractional neuro-sequential ARFIMA-LSTM for financial market forecasting. *IEEE Access* 8: 71326–38. [CrossRef]

Chou, Chih-Hsun, Su-Chen Hsieh, and Chui-Jie Qiu. 2017. Hybrid genetic algorithm and fuzzy clustering for bankruptcy prediction. *Applied Soft Computing* 56: 298–316. [CrossRef]

Cleofas-Sánchez, Laura, Vicente. García Jiménez, Ana Isabel Marqués, and J. S. Sánchez. 2016. Financial distress prediction using the hybrid associative memory with translation. *Applied Soft Computing* 44: 144–52. [CrossRef]

Dourado, A. 2013. How should I choose the optimum number for the neurons in the input/hidden layer for a recurrent neural network?

El-Bannany, Magdi, Ahlam H. Dehghan, and Ahmed M. Khedr. 2021. Prediction of Financial Statement Fraud using Machine Learning Techniques in UAE. Paper presented at the 2021 18th International Multi-Conference on Systems, Signals & Devices (SSD), Monastir, Tunisia, March 22–25; pp. 649–54.

El-Bannany, Magdi, Meenu Sreedharan, and Ahmed M. Khedr. 2020. A robust deep learning model for financial distress prediction. *International Journal of Advanced Computer Science and Applications (IJACSA)* 11: 170–75. [CrossRef]

El Bannany, Magdi, Ahmed M. Khedr, Meenu Sreedharan, and Sakeena Kanakkayil. 2021. Financial Distress Prediction based on Multi-Layer Perceptron with Parameter Optimization. *IAENG International Journal of Computer Science* 48: 1–12.

Falbo, Paolo. 1991. Credit-scoring by enlarged discriminant models. *Omega* 19: 275–89. [CrossRef]

Geng, Ruibin, Indranil Bose, and Xi Chen. 2015. Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research* 241: 236–47. [CrossRef]

Gepp, Adrian, and Kuldeep Kumar. 2015. Predicting Financial Distress: A Comparison of Survival Analysis and Decision Tree Techniques. *Procedia Computer Science* 54: 396–404. [CrossRef]

Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. Paper presented at the 28th International Conference on Machine Learning (ICML-11), Bellevue, WA, USA, 28 June–2 July.

Halim, Zulkifli, Shuhaida Mohamed Shuhidan, and Zuraidah Mohd Sanusi. 2021. Corporation financial distress prediction with deep learning: Analysis of public listed companies in Malaysia. *Business Process Management Journal* 27: 1163–78. [CrossRef]

Hu, Hui, and Milind Sathye. 2015. Predicting Financial Distress in the Hong Kong Growth Enterprises Market from the Perspective of Financial Sustainability. *Sustainability* 7: 1186–1200. [CrossRef]

Huang, Yu-Pei, and Meng-Feng Yen. 2019. A new perspective of performance comparison among machine learning algorithms for financial distress prediction. *Applied Soft Computing* 83: 105663. [CrossRef]

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9: 1735–80. [CrossRef] [PubMed]

Hou, Baozhen. 2016. Financial distress prediction of k-means clustering based on genetic algorithm and rough set theory. *Chemical Engineering Transactions* 51: 505–10.

Jang, Youjin, Inbae Jeong, Yong K. Cho, and Yong Han Ahn. 2019. Predicting business failure of construction contractors using long short-term memory recurrent neural network. *Journal of Construction Engineering and Management* 145: 04019067. [CrossRef]

Khan, Naushad, Shah Fahad, Mahnoor Naushad, and Shah Faisal. 2020. COVID-2019 Locked Down Effects on Oil Prices and Its Effects on the World Economy. *Available at SSRN 3588810*. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3588810 (accessed on 10 January 2023).

Khedr, Ahmed M., Ifra Arif, Magdi El-Bannany, Saadat M. Alhashmi, and Meenu Sreedharan. 2021. Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey. *Intelligent Systems in Accounting, Finance and Management* 28: 3–34. [CrossRef]

Khedr, Ahmed M., Magdi El Bannany, and Sakeena Kanakkayil. 2021. An ensemble model for financial statement fraud detection. *ARPHA Preprints* 2: e69590.

Kim, Kyoung-jae, Kichun Lee, and Hyunchul Ahn. 2019. Predicting Corporate Financial Sustainability Using Novel Business Analytics. *Sustainability* 11: 64. [CrossRef]

Matin, Rastin, Casper Hansen, Christian Hansen, and Pia Mølgaard. 2019. Predicting distresses using deep learning of text segments in annual reports. *Expert Systems with Applications* 132: 199–208. [CrossRef]

Mehreen, Mehreen, Maran Marimuthu, Samsul Ariffin Abdul Karim, and Amin Jan. 2020. Proposing a multidimensional bankruptcy prediction model: An approach for sustainable Islamic banking. *Sustainability* 12: 3226. [CrossRef]

Min, Jae H., and Young-Chan Lee. 2005. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications* 28: 603–14. [CrossRef]

Ohlson, James A. 1980. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research* 18: 109–31. [CrossRef]

Ravisankar, P., and V. Ravi. 2010. Financial distress prediction in banks using Group Method of Data Handling neural network, counter propagation neural network and fuzzy ARTMAP. *Knowledge-Based Systems* 23: 823–31. [CrossRef]

Ribeiro, Bernardete, and Noel Lopes. 2011. Deep belief networks for financial prediction. Paper presented at the International Conference on Neural Information Processing, Shanghai, China, November 13–17. Berlin and Heidelberg: Springer, pp. 766–73.

Sun, Jie, Hamido Fujita, Peng Chen, and Hui Li. 2017. Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vector machine ensemble. *Knowledge-Based Systems* 120: 4–14. [CrossRef]

Sun, Jie, and Xiaofeng Hui. 2006. An application of decision tree and genetic algorithms for financial ratios' dynamic selection and financial distress prediction. Paper presented at the 2006 International Conference on Machine Learning and Cybernetics, Dalian, China, August 13–16; pp. 2413–18.

Shen, Furao, Jing Chao, and Jinxi Zhao. 2015. Forecasting exchange rate using deep belief networks and conjugate gradient method. *Neurocomputing* 167: 243–253. [CrossRef]

Shin, Kyung-Shik, and Yong-Joo Lee. 2002. A genetic algorithm application in bankruptcy prediction modeling. *Expert Systems with Applications* 23: 321–28. [CrossRef]

Sreedharan, Meenu, Ahmed M. Khedr, and Magdi El Bannany. 2020a. A multi-layer perceptron approach to financial distress prediction with genetic algorithm. *Automatic Control and Computer Sciences* 54: 475–482. [CrossRef]

Sreedharan, Meenu, Ahmed M. Khedr, and Magdi El Bannany. 2020b. A comparative analysis of machine learning classifiers and ensemble techniques in financial distress prediction. Paper presented at the 2020 17th International Multi-Conference on Systems, Signals & Devices (SSD), Monastir, Tunisia, 20–23 July; pp. 653–57.

Wanke, Peter, Carlos Barros, and João Faria. 2015. Financial distress drivers in Brazilian banks: A dynamic slacks approach. *European Journal of Operational Research* 240: 258–68. [CrossRef]

Wu, Desheng, Liang Liang, and Zijiang Yang. 2008. Analyzing the financial distress of Chinese public companies using probabilistic neural networks and multivariate discriminate analysis. *Socio-Economic Planning Sciences* 42: 206–20. [CrossRef]

Zhu, Lin, Dawen Yan, Zhihua Zhang, and Guotai Chi. 2022. Financial Distress Prediction of Chinese Listed Companies Using the Combination of Optimization Model and Convolutional Neural Network. *Mathematical Problems in Engineering* 2022: 9038992. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

sustainability

MDPI

Article

Comparing the Performance of Corporate Bankruptcy Prediction Models Based on Imbalanced Financial Data

Seol-Hyun Noh

Department of Statistical Data Science, ICT Convergence Engineering, Anyang University, Anyang 14028, Republic of Korea; shnoh@anyang.ac.kr

Abstract: Forecasts of corporate defaults are used in various fields across the economy. Several recent studies attempt to forecast corporate bankruptcy using various machine learning techniques. We collected financial information on 13 variables of 1020 companies listed on the KOSPI and KOSDAQ to capture the possibility of corporate bankruptcy. We propose a data processing method for small-sample domestic corporate financial data. We investigate the case of random sampling of non-bankrupt companies versus sampling non-bankrupt companies based on approximate entropy and optimized threshold based on AUC to address the imbalance between the number of bankrupt companies and the number of non-bankrupt companies. We compare the performance measures of corporate bankruptcy prediction models for the small sample data structured in two ways and the full dataset. The experimental results of this study contribute to the selection of an appropriate corporate bankruptcy prediction model.

Keywords: corporate bankruptcy; bankruptcy prediction; performance comparison; imbalanced financial data

1. Introduction

Many sectors of the economy use predictions of corporate bankruptcy. A bankruptcy prediction model allows companies to diagnose the current state and establish strategies and management to operate the business more stably by managing key indicators that affect its bankruptcy risk. It also allows investors to establish investment strategies and improve their portfolios and governments to establish policies to improve macroeconomic soundness and financial regulatory policies.

Several recent studies use various methods to predict corporate bankruptcy. In particular, since the 1990s, computational economics focused on predicting corporate bankruptcy with artificial neural networks [1–10].

Barboza et al. [3] compare machine learning models with traditional models for bankruptcy prediction and show that machine learning models had 10% better performance in accuracy on average. Specific machine learning bankruptcy prediction models include classification models such as support vector machines (SVM), logistic regression (LR), k-nearest neighbors (k-NN), decision tree (DT), and random forest (RF), and time series models such as recurrent neural network (RNN) and long short-term memory (LSTM). These models are widely used as consumer bankruptcy prediction models and corporate bankruptcy prediction models [11–14].

Classification is a supervised learning technique to conduct predictive analytics with a categorical outcome, whether binary or multiclass. Much current research concentrates on classification using several algorithms from basic to advanced, such as logistic regression, discriminant analysis, Naïve Bayes, decision tree, random forest, support vector machine, neural network, and so on [15]. These are well developed and successfully applied to many application domains. However, the majority of supervised learning techniques were developed for balanced class distribution, leaving imbalanced class distribution relatively

neglected. For this reason, many researchers seek to address the imbalance between the number of bankrupt companies and the number of non-bankrupt companies.

Kim et al. [7] use RNN and LSTM models to predict bankruptcy and compare their performance with other classification models, including of SVM, LR, and RF. From experiments, Kim et al. [7] demonstrate that the RNN and LSTM models outperformed in terms of accuracy, precision, recall, and AUC, using the monthly financials from the Compustat North America Dataset for about 45,472 non-financial firms for January 2007 to December 2019. Since 2057 of the firms went bankrupt, the authors had a large amount of monthly financial information on the bankrupt companies and thus had a sufficient training dataset for the financial information of the bankrupt companies for time series models such as RNN and LSTM. As a balanced classification, accuracy may be the unbiased metric for evaluation. Therefore, a corporate bankruptcy prediction model suitable for unbalanced corporate financial data is necessary.

In this study, we propose a random small sampling method of non-bankrupt companies and a method of small sampling of companies representing non-bankrupt companies based on approximate entropy to improve corporate bankruptcy prediction performance for imbalanced corporate financial data. To deal with the imbalanced class, we also propose an optimal threshold method to improve AUC performance even when considering the imbalanced total dataset. The optimal threshold method proposed in this study applies the concept of the anomaly score defined in [16] to corporate bankruptcy prediction models. We provide experimental results to analyze the effectiveness of models for predicting corporate bankruptcy for imbalanced corporate financial information in Section 3.

To derive these results, we established the following research questions:

- Research question 1: How can we derive data sampling methods that improve the performance of corporate bankruptcy prediction models for imbalanced corporate financial information?
- Research question 2: How can we derive an optimal threshold technique that improve AUC performance even when considering the imbalanced corporate financial information?

In order to solve the above two research questions, the research results were derived according to the research process as shown in Figure 1.

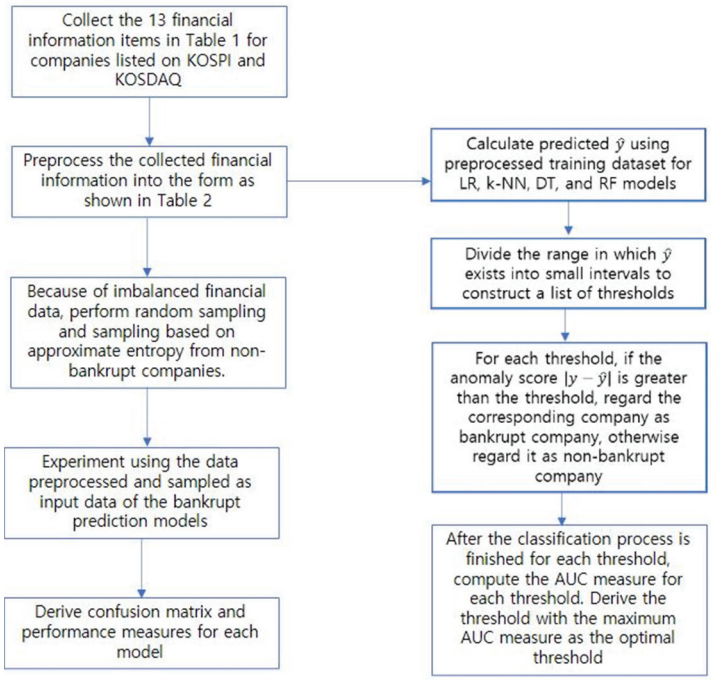


Figure 1. Diagram of research stage.

2. Data and Methods

2.1. Data and Sampling

2.1.1. Data

We collected the 13 financial information items for about 1020 companies listed on the Korea Composite Stock Price Index (KOSPI) and Korean Securities Dealers Automated Quotations (KOSDAQ) from 2012 to 2021. As 20 companies went bankrupt, it is not possible to build a sufficient training dataset for the RNN and LSTM time series models. Table 1 summarizes the items collected for this sample, excluding the parent company equity holder item. The corporate financial information in Table 1 corresponds to the criteria that characterize bankrupt and non-bankrupt companies [2]. Cha and Kang [2] conducted a multivariate discriminant analysis to select the 29 variables like in Appendix A that correspond to the criteria characterizing bankrupt and non-bankrupt companies from a set of continuous independent variables of financial information. After calculating the correlation coefficient of these 29 variables to create highly correlated variable groups, we conducted a t-test to detect statistically significant differences between the population mean of bankrupt and non-bankrupt companies within each variable group. Table 1 presents the resulting 14 variables with the lowest p-values. We collected the 13 financial information items from FnGuide through the FnDB Navigator.

Table 1. Financial information collected for each sample company (excluding the parent company equity holder item. Source: [2]).

Financial Information	*t*-Test *p*-Value
Total assets	2.33×10^{-5}
Parent company equity holder	3.59×10^{-4}
Intangible assets ratio	4.84×10^{-2}
Equity capital ratio	1.48×10^{-20}
Debt ratio	9.63×10^{-6}
Cash flows/total liabilities	2.23×10^{-8}
Total assets growth rate	2.88×10^{-14}
Operating revenue/operating expense	6.71×10^{-13}
Gross margin	2.45×10^{-4}
ROA (income from continuing operations before tax)	1.79×10^{-9}
ROA (operating profit)	2.06×10^{-10}
ROE (income from continuing operations before tax)	2.89×10^{-5}
ROE (operating profit)	2.52×10^{-2}
Total debt turnover ratio	5.68×10^{-27}

Among the 1020 listed companies, we define those delisted between 2012 and 2021 as bankrupt and used the financial data of the 20 bankrupt companies and 1000 non-bankrupt companies as the input for the LSTM, LR, k-NN, DT, and RF bankruptcy prediction models. To test bankruptcy prediction performance while preventing overfitting in the models caused by the data of non-bankrupt companies due to the small number of bankrupt companies, we conducted an experiment using the data of all companies, a subsample of 10 bankrupt companies, and a subsample of 20 non-bankrupt companies for the training dataset and test dataset. Table 2 summarizes the collected financial information configured as the input and label data of the model. Figure 2 shows the input data and label data of LR, k-NN, DT, and RF bankruptcy prediction models. For companies that went bankrupt in 2013, financial data for 2013 were padded and used as financial data from 2014 to 2021, and for companies that went bankrupt in 2020, financial data for 2020 were used as financial data for 2021. We applied the same padding technique to label data [17].

Table 2. Input and label data in the proposed bankruptcy prediction model.

LSTM input data: Company (Financial information in 2012, . . . , 2020)
LSTM label data: Company (Bankruptcy or non-bankruptcy): (1)/(0)
LR, k-NN, DT, RF input data: Company (Financial information in 2012) ⋮ (Financial information in 2020)
LR, k-NN, DT, RF label data: Company (Bankruptcy or non-bankruptcy in 2013) ⋮ (Bankruptcy or non-bankruptcy in 2021)

node_id	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	time
924	49.73	64.10	−5.43	18.37	29.08	110.42	8.32	−6.08	16.32	−11.94	1.80	38,179,277	6,062,267	2012
984	67.94	18.59	−15.54	23.56	19.92	93.37	−3.04	−7.07	−3.94	−9.14	1.89	25,364,197	3,365,951	2012
⋮														⋮
511	39.17	108.57	15.10	96.21	30.45	106.42	7.99	3.48	18.53	8.06	2.33	16,634,320	526,704	2012
512	73.36	12.03	14.68	10.94	5.96	99.40	−0.90	2.22	−1.18	2.89	6.49	149,560,052	1895	2012
⋮														⋮

(a) Input data of LR, k-NN, DT, and RF bankruptcy prediction models

node_id	2013	2014	2015	2016	2017	2018	2019	2020	2021
924	0	0	0	1	1	1	1	1	1
984	0	1	1	1	1	1	1	1	1
⋮									
511	0	0	0	0	0	0	0	0	0
512	0	0	0	0	0	0	0	0	0
⋮									

(b) Label data of LR, k-NN, DT, and RF bankruptcy prediction models

Figure 2. Input data and label data of the models. In panel (**a**), v1 to v13 refer to the 13 financial information items.

As we must configure the input data for the LSTM model with time series data in an experiment on all of the 1020 companies, the training and testing datasets were configured with data from 10 bankrupt companies and 500 non-bankrupt companies for the experiment. For the LR, k-NN, DT, and RF models, we set the financial information in the n-th year as the input data and bankruptcy or non-bankruptcy in year $n + 1$ as the label data, creating 9 times more input data than the LSTM model for each company. Hence, the training and testing datasets were configured with the data of 90 bankrupt companies and 4500 non-bankrupt companies for the experiment. As the input data for the LR, k-NN, DT, and RF models have 9 times as many bankrupt companies as the LSTM model, we expect better bankruptcy prediction performance.

With data from a small sample of bankrupt companies, overfitting occurs for non-bankrupt companies in the LR, k-NN, DT, and RF models when conducting an experiment using all companies. As a result, bankruptcy prediction performance of the models decreases. Therefore, considering the number of bankrupt companies, we set the financial information of 10 bankrupt companies and 20 non-bankrupt companies as small sample data consisting of the training and testing datasets and conducted a comparison with an experiment using all 1020 companies. For a small sample, we randomly sampled the non-bankrupt companies 5 times and report the mean performance measure of the 5 repeated experiments.

Similarly, we set the training and testing datasets for the LSTM model using the data of 10 bankrupt companies and 20 non-bankrupt companies for the experiment.

2.1.2. Sampling

In the small-sample data experiment, we used the financial information of bankrupt 10 companies and 20 non-bankrupt companies as the training dataset. We found performance differences depending on the sampling method used to select the 20 non-bankrupt companies from the 1000 non-bankrupt companies, as described in Section 3. To reduce this difference, we conducted random sampling of the non-bankrupt companies 5 times and used the mean performance measure of 5 repeated experiments. As data imbalances may still occur using this method, we set the time series for each of the 13 financial information items and compared the performance with the sampling of 20 non-bankrupt companies which we consider as having characteristics that represent non-bankrupt companies as their entropy is low.

The approximate entropy we used is a method that measures the entropy of a time series, with the following definition [18]: we configure $N - m + 1$ vectors $x_m(i) = [u(i + k) : 0 \leq k \leq m - 1]$ for time series $[u(j) : 1 \leq j \leq N]$. The number of vectors $x_m(j)$ within the distance of r from $x_m(i)$ is set as B_i and the number of vectors $x_{m+1}(j)$ within the distance of r from $x_{m+1}(i)$ as A_i. m is the length of compared sequences, and r is a matching tolerance, where the constant is determined before measuring approximate entropy. The distance between two vectors $x_m(i)$ and $x_m(j)$ is

$$d(x_m(i), x_m(j)) = max\{|u(i + k) - u(j + k)| : 0 \leq k \leq m - 1\},$$

where $C_i^m(r) = \frac{B_i}{N-m+1}$, $C_i^m(r)$ is the probability that vector $x_m(j)$ exists within the distance of r with $x_m(i)$.

Approximate entropy $ApEn(m, r, N)$ is

$$\frac{1}{N-m+1}\sum_{i=1}^{N-m+1} ln|C_i^m(r)| - \frac{1}{N-m}\sum_{i=1}^{N-m} ln\left|C_i^{m+1}(r)\right|,$$

which is approximately $\frac{1}{N-m}\sum_{i=1}^{N-m}\left\{-ln\left(\frac{A_i}{B_i}\right)\right\}$.

Companies with a low approximate entropy across financial information items are considered to represent the characteristics of non-bankrupt companies, as they have a low degree of disorder across items. We sampled 20 non-bankrupt companies in the training dataset in descending order of approximate entropy and compared the results with an experiment using the data on a random, small sample. The mean approximate entropy was measured for the time series of each financial information item, and the approximate entropy of 20 sampled non-bankrupt companies was within 1.70. To measure approximate entropy $m = 3$, we set r as the standard deviation of each financial information item.

2.2. Models and Performance Measures

We tested the LSTM, LR, k-NN, DT, and RF models as corporate bankruptcy prediction models. The LR, k-NN, DT, and RF models used the Keras Library, and the LSTM model was implemented using Python 3.8 for Ubuntu Linux as follows.

2.2.1. LSTM Model

The LSTM model is a deep learning model that uses time-series data as the input and outputs future data, and it is used to predict river levels, solar power generation, fine dust, energy demand, and stock prices [19]. Using the dataset in Section 2.1.1 as the input and label data, we establish a model to predict bankruptcy in the next year. The mathematical model of the LSTM is expressed as in Equation (1) and Figure 3. Output h_t, outputgate o_t, memory cell c_t, new memory content $\widetilde{c_t}$, forget gate f_t, and input gate i_t in the LSTM are modeled as in Equation (1) and illustrated in Figure 2. The sigmoid function $\sigma(x)$ is

defined as $\sigma(x) = \frac{1}{1+e^{-x}}$ for real value x. If X is a matrix with real numbers as elements, $\sigma(X)$ is a matrix with the values of the sigmoid function for each element of X as elements

LSTM model pseudocode
Step (1) train_corp_data = 9-year data of 13 financial information items of 10 bankrupt companies test_corp_data = 9-year data of 13 financial information items of the other 10 bankrupt companies train_corp_label = whether a company went bankrupt during 6 years, obtained by setting 4 years as the sequence length for train_corp_data test_corp_label = whether a company went bankrupt during 6 years, obtained by setting 4 years as the sequence length for test_corp_data Step (2) Normalizing 4 datasets in Step 1 Step (3) Stacking 4 years of financial information corresponding to the sequence length as input data for each LSTM Step (4) Setting the hidden size of the LSTM at 256 and building the LSTM model using nn.LSTM() provided from torch.nn Step (5) Store the output value of the LSTM in the output variable, while forwarding the input data of the LSTM through the device. Step (6) Calculating a performance measure by outputting the confusion matrix

$$\text{Output } h_t = o_t \; tanh(c_t) \tag{1}$$

$$\text{Output gate } o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_t)$$
$$\text{Memory cell } c_t = f_t c_{t-1} + i_t \widetilde{c}_t$$
$$\text{New memory content } \widetilde{c}_t = tanh(W_c x_t + U_c h_{t-1})$$
$$\text{Forget gate } f_t = \sigma\left(W_f x_t + U_f h_{t-1} + V_f c_{t-1}\right)$$
$$\text{Input gate } i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i c_{t-1})$$

2.2.2. Logistic Regression Model

The logistic regression model is a method of modeling the conditional probability $P(Y = 1|x_1, x_2, \cdots, x_k)$ as shown in Equation (2), where $X_1,\ X_2,\ \cdots,\ X_k$ are explanatory variables and Y (Y is 0 or 1) is the binary response variable.

$$log\left[\frac{P(Y = 1|x_1, x_2, \cdots, x_k)}{1 - P(Y = 1|x_1, x_2, \cdots, x_k)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{2}$$

From the training dataset, intercept α and the effect of x_i, namely β_i $(i = 1,\ 2,\ \cdots, k)$, are estimated with maximum likelihood estimation.

It is an algorithm that solves the problem of binary classification as it considers new data as $Y = 1$ if $P(Y = 1|x_1, x_2, \cdots, x_k)$ obtained from Prediction Equation (2) is greater than the pre-defined threshold and $Y = 0$ if it is smaller than the threshold.

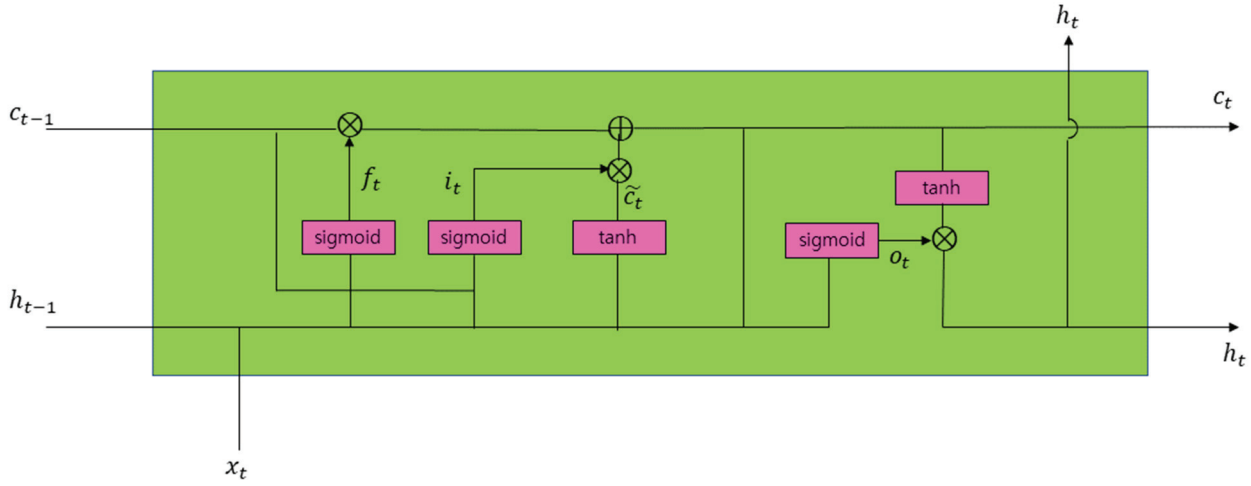


Figure 3. LSTM model [19].

2.2.3. k-Nearest Neighbor (k-NN) Model

The k-nearest neighbor algorithm is a non-parametric supervised learning method as a machine learning algorithm. The k-NN is a method of extracting the nearest k using the distance measuring metric d among the existing data when new data are given and predicting the class of new data for a classification problem and the prediction value of new data for a regression problem based on the information of the extracted data. The hyperparameters of the k-NN model include the number of searchable neighbors k and metric d. The Euclidean distance, Manhattan distance, Mahalanobis distance, correlation distance, and rank correlation distance are used as d.

2.2.4. Decision Tree Model

The decision tree model is a supervised machine learning method to solve regression or classification problems and conduct prediction by categorizing decision-making rules into a tree structure. The decision tree model is created using learning data and consists of a hierarchy of branches in which the explanatory variables are expressed as nodes and feature spaces are categorized into non-overlapping groups based on certain conditions. Each internal node represents a test on an attribute (e.g., whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represents classification rules. In the final nodes at the bottom, the dependent variables are shown as the categorized groups.

2.2.5. Random Forest Model

The random forest (RF) model is an ensemble learning method as a machine learning algorithm using multiple DTs. The RF algorithm is a method for solving regression and classification problems. For classification problems, the class predicted by most DTs is produced as the output. For regression problems, the average of the predicted values of each DT is produced as the output. The RF algorithm is designed to solve the problem of overfitting in the training dataset for the DT. It divides the training dataset into different parts and randomly selects the pre-defined number of explanatory variables for each of the training dataset parts to generate a new DT.

This method reduces variance in the model by learning several DTs and averaging the prediction values from the DTs. Figure 4 is a diagram illustrating the RF algorithm.

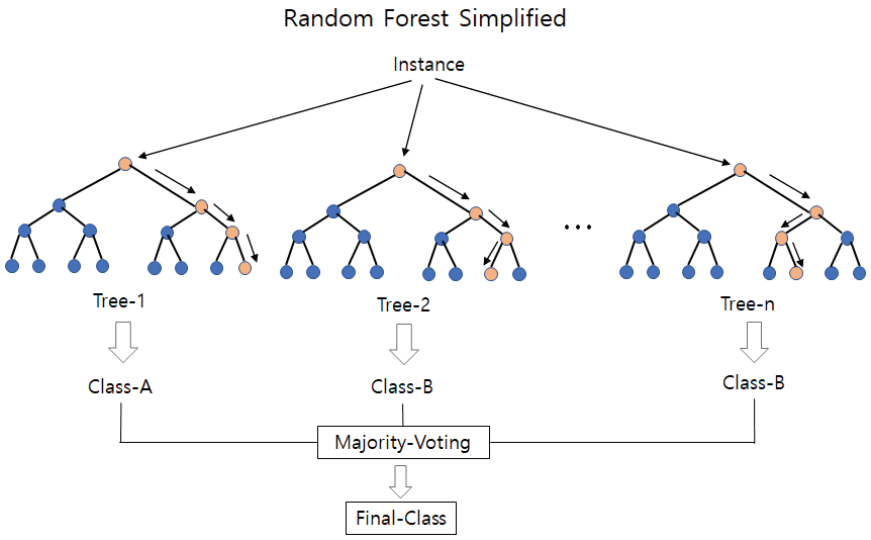


Figure 4. Diagram of random forest (source: [20]).

2.2.6. Performance Measure

As a small sample is not considered to be imbalanced data, we categorize a bankrupt company if the probability value calculated for the test dataset from 5 trained models is greater than the threshold of 0.5 and as a non-bankrupt company if it is not. As an actual non-bankrupt company can be predicted as non-bankrupt or bankrupt and the actual bankrupt company can be predicted as non-bankrupt or bankrupt, this results in the confusion matrix in Table 3.

Table 3. Confusion matrix.

		Prediction outcome	
		Non-bankrupt company	Bankrupt company
Actual outcome	Non-bankrupt company	True positive (TP)	False negative (FN)
	Bankrupt company	False positive (FP)	True negative (TN)

Accuracy, non-bankruptcy precision, non-bankruptcy recall, bankruptcy precision, bankruptcy recall, non-bankruptcy F1 score, and bankruptcy F1 score can be calculated based on the confusion matrix. By comparing the experimental results and the actual label values, we can create the graph of the ROC curve and the area under the ROC curve (AUC). The definition of each measure is provided in Equation (3).

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+FP+TN+FN} \\ Non\text{-}bankruptcy\ Precision &= \frac{TP}{TP+FP} \\ Non\text{-}bankruptcy\ Recall &= \frac{TP}{TP+FN} \\ Bankruptcy\ Precision &= \frac{TN}{FN+TN} \\ Bankruptcy\ Recall &= \frac{TN}{FP+TN} \end{aligned} \tag{3}$$

$$Non\text{-}bankruptcy\ F1\ Score = \frac{2 \times Non\text{-}bankruptcy\ Precision \times Non\text{-}bankruptcy\ Recall}{Non\text{-}bankruptcy\ Precision + Non\text{-}bankruptcy\ Recall}$$

$$Bankruptcy\ F1\ Score = \frac{2 \times Bankruptcy\ Precision \times Bankruptcy\ Recall}{Bankruptcy\ Precision + Bankruptcy\ Recall}$$

The ROC AUC graph has the false positive rate as the x-axis and the true positive rate as the y-axis. It is drawn above the baseline $y = x$ and convex upward. A graph drawn above the baseline means better classification performance in the model. Hence, a higher AUC value means the model has higher classification performance.

3. Performance Analysis Results

Table 4 shows the performance results in terms of the confusion matrix, accuracy, non-bankruptcy precision, non-bankruptcy recall, bankruptcy precision, bankruptcy recall, non-bankruptcy F1 score, bankruptcy F1 score, and AUC for the LR, k-NN, DT, and RF models. The analysis was conducted for the small sample dataset created based on the 5 random samplings of 20 non-bankrupt companies, for the small sample dataset with 20 non-bankrupt companies chosen based on approximate entropy, and the total data.

The analysis of the LSTM model for the small sample data returned a confusion matrix of $\begin{bmatrix} 20 & 0 \\ 10 & 0 \end{bmatrix}$ and the model did not learn the characteristics of bankrupt companies very well as the sample of 10 companies is small. The analysis of the LSTM model for the total data returned a confusion matrix of $\begin{bmatrix} 500 & 0 \\ 10 & 0 \end{bmatrix}$ and the model did not learn the characteristics of bankrupt companies very well given the small number of 10 bankrupt companies, much like the small sample data.

Panel (a) in Table 4 shows the performance results of the models with the small dataset sampled randomly five times. The k-NN had the largest AUC, at 0.8676, and the RF had the second largest AUC, at 0.8095, indicating that the k-NN and RF outperformed the LR and DT in diagnosing non-bankrupt companies.

The LR had the highest non-bankruptcy F1 score in the performance measures of non-bankruptcy prediction and non-bankruptcy recall, while the DT had the highest bankruptcy F1 score in the performance measures of bankruptcy precision and bankruptcy recall. The LR is the best among the five models in non-bankruptcy prediction and non-bankruptcy recall, while the DT is the best in bankruptcy prediction and bankruptcy recall.

Panel (b) in Table 4 provides the performance measure results of the models for the small sample data sampled using approximate entropy. Here, RF had the highest AUC, at 0.8326, and the k-NN had the second highest AUC, at 0.8076. Further, RF and the k-NN outperformed the LR and DT in diagnosing non-bankrupt companies. RF had the best performance in terms of non-bankruptcy precision, non-bankruptcy recall, bankruptcy precision, and bankruptcy recall. Thus, RF is the best of the five models in terms of precision and recall.

The performance results in Panels (a) and (b) in Table 4 show that the AUC of the k-NN and RF was higher than that of the LR and DT. The analysis results for the small dataset sampled using approximate entropy showed better performance measures of bankruptcy precision and bankruptcy recall than those for the small sample data sampled randomly.

Panel (c) in Table 4 shows the performance results of the models for the total data. Like the LSTM, the small number of 10 bankrupt companies led to overfitting to non-bankrupt companies and subsequently resulted in 0 or close to 0 in bankruptcy precision and bankruptcy recall, despite generating samples for 90 bankrupt companies, which is 9 times larger than the input data. These models did not predict bankruptcies very well.

Table 4. Performance results: LR, k-NN, DT, and RF models for the small sample and total dataset.

Model	Confusion Matrix	Accuracy	Non-Bankruptcy Precision	Non-Bankruptcy Recall	Bankruptcy Precision	Bankruptcy Recall	Non-Bankruptcy F1 Score	Bankruptcy F1 Score	AUC
LR	$\begin{bmatrix} 183 & 23 \\ 49 & 15 \end{bmatrix}$	0.7333	0.7888	0.8883	0.3947	0.2344	0.8356	0.2941	0.7457
k-NN	$\begin{bmatrix} 183 & 23 \\ 53 & 11 \end{bmatrix}$	0.7185	0.7754	0.8883	0.3235	0.1719	0.8281	0.2245	0.8676
DT	$\begin{bmatrix} 155 & 51 \\ 42 & 22 \end{bmatrix}$	0.6556	0.7868	0.7524	0.3014	0.3438	0.7692	0.3212	0.6587
RF	$\begin{bmatrix} 178 & 28 \\ 49 & 15 \end{bmatrix}$	0.7148	0.7841	0.8641	0.3488	0.2344	0.8222	0.2804	0.8095

(a) Performance results of the LR, k-NN, DT, and RF for the small random sample

Model	Confusion Matrix	Accuracy	Non-Bankruptcy Precision	Non-Bankruptcy Recall	Bankruptcy Precision	Bankruptcy Recall	Non-Bankruptcy F1 Score	Bankruptcy F1 Score	AUC
LR	$\begin{bmatrix} 187 & 9 \\ 58 & 16 \end{bmatrix}$	0.7519	0.7633	0.9541	0.64	0.2162	0.8481	0.3232	0.7929
k-NN	$\begin{bmatrix} 187 & 9 \\ 53 & 21 \end{bmatrix}$	0.7704	0.7792	0.9541	0.7	0.2838	0.8578	0.4038	0.8076
DT	$\begin{bmatrix} 168 & 28 \\ 41 & 33 \end{bmatrix}$	0.7444	0.8038	0.8571	0.5410	0.4459	0.8296	0.4889	0.5979
RF	$\begin{bmatrix} 182 & 14 \\ 43 & 31 \end{bmatrix}$	0.7889	0.8089	0.9286	0.6889	0.4189	0.8646	0.5210	0.8326

(b) Performance results: LR, k-NN, DT, and RF for the small dataset sampled using approximate entropy

Model	Confusion Matrix	Accuracy	Non-Bankruptcy Precision	Non-Bankruptcy Recall	Bankruptcy Precision	Bankruptcy Recall	Non-Bankruptcy F1 Score	Bankruptcy F1 Score	AUC_1 / AUC_2
LR	$\begin{bmatrix} 4514 & 2 \\ 74 & 0 \end{bmatrix}$	0.9834	0.9839	0.9995	0	0	0.9916	Not defined	0.7906 / 0.9998
k-NN	$\begin{bmatrix} 4503 & 13 \\ 74 & 0 \end{bmatrix}$	0.9810	0.9839	0.9971	0	0	0.9904	Not defined	0.6500 / 0.9998
DT	$\begin{bmatrix} 4398 & 118 \\ 68 & 6 \end{bmatrix}$	0.9595	0.9848	0.9739	0.0484	0.0811	0.9793	0.0606	0.5477 / 0.9998
RF	$\begin{bmatrix} 4514 & 2 \\ 74 & 0 \end{bmatrix}$	0.9834	0.9839	0.9996	0	0	0.9917	Not defined	0.7389 / 0.9998

Table 4. *Cont.*

(c) Performance results: LR, k-NN, DT, and RF for the total data, where AUC_1 = AUC at the threshold of 0.5 and AUC_2 = AUC using the optimal threshold.					
	Accuracy	**Precision**	**Recall**	**F1 Score**	**AUC**
Logistic	0.7570	0.0001	0.2174	0.0002	0.4872
SVM	0.7236	0.0002	0.3913	0.0003	0.5575
RF	0.9899	0.0023	0.2174	0.0045	0.6037
RNN	0.9789	0.0024	0.4783	0.0048	0.7286
LSTM	0.9936	0.0058	0.3478	0.0114	0.6707
Ensemble	0.9826	0.0029	0.4783	0.0058	0.7305
(d) Performance results: Bankruptcy Forecasting Performance by Methodology [7]					

To overcome the problem of worsening performance due to data imbalance, we propose a method to improve the AUC performance measure by applying the optimal threshold. The LR, k-NN, DT, and RF models predict bankruptcy in the $(n + 1)$-th year based on n-th year's financial information. The anomaly score is the difference between the value indicating bankruptcy by the model in year $(n + 1)$ and the value indicating actual bankruptcy (bankruptcy is 1, non-bankruptcy is 0). A higher anomaly score means the status in year $(n + 1)$ is likely to be an anomaly (bankruptcy). We use a supervised thresholding mechanism to detect the optimal threshold for binarizing points as anomalous or not. We obtain the minimum and maximum values of the anomaly scores calculated on the training dataset and then consider each value from the minimum score to the maximum score with a small step and select the value returning the highest F-score as the overall model threshold. Figure 5 shows the optimal threshold mechanism.

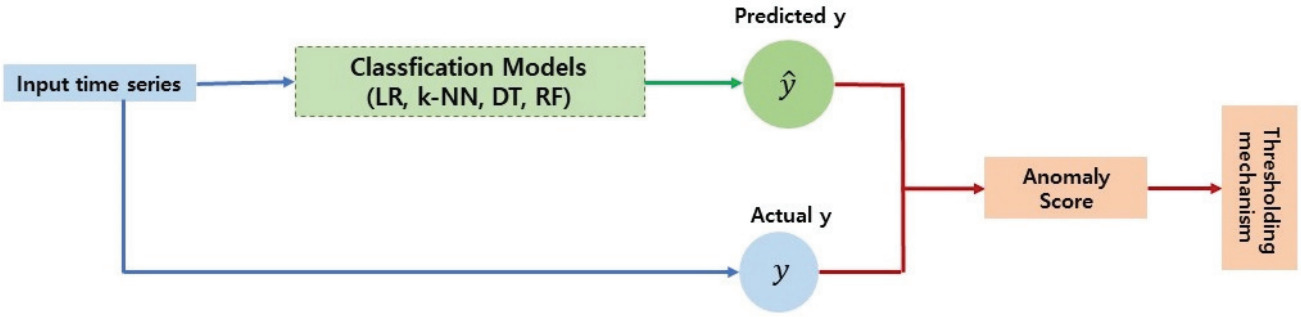


Figure 5. Optimal threshold mechanism.

We believe that our proposed method would greatly enhance the value of the AUC as in Panel (c) in Table 4, where the optimal threshold is applied. Using the information obtained from the anomaly scores from the training dataset, we use the AUC-based thresholds mechanism to identify the optimal thresholds for determining bankruptcy or non-bankruptcy. The anomaly score is measured for the test dataset using the identified optimal thresholds. If the anomaly score is above the identified threshold, then we regard it as a bankruptcy (1). If it is below the identified threshold, then we regard it as non-bankrupt (0).

As shown in Table 4, the AUC of the performance measures in Panels (a) and (b) was generally larger than that in Panel (c), and the performance measures of bankruptcy precision and bankruptcy recall were significant, suggesting that the method of learning and predicting bankruptcy by configuring only some non-bankrupt companies as the training dataset when there are few bankrupt companies can create a model that classifies bankrupt and non-bankrupt companies well without greatly sacrificing accuracy. However, the AUC value when the optimal threshold is applied in Panel (c) in Table 4 is greater than the AUC of the performance measures in Panels (a) and (b) in Table 4, which suggests that identifying the optimal threshold based on the AUC for classification is more effective than applying small sampling to imbalanced data.

The observed difference in performance according to the models suggests that the RF is the best model, with the highest AUC of 0.8326 in Panel (b) in Table 4 to diagnose non-bankrupt companies. It is also the best model, with the highest non-bankruptcy F1 and bankruptcy F1 scores of 0.8646 and 0.5210, respectively, in Panel (b), to classify non-bankrupt and bankrupt companies. Panel (d) in Table 4 is the result of [7], an analysis of monthly financial information of about 45,472 non-financial firms collected from January 2007 to December 2019 from the Compustat North America Dataset. Since it included the monthly financial information of 2057 bankrupt companies, it offers a sufficient training dataset for the RNN and LSTM. Hence, the RNN and LSTM had significant AUC performance.

The performance measures of precision, recall, F1 score, and AUC from the LR and RF models trained with the small random sample and the LR and RF models trained with the small sample created using approximate entropy were higher than those of the LR and RF models in Panel (d) in Table 4. In addition, the AUC performance measure in Panel (c) and the AUC performance measure when the optimal threshold was applied were higher than those of the LR and RF models in Panel (d).

These results suggest that we identified a data processing method and two data sampling methods that improve the performance of the corporate bankruptcy prediction models for imbalanced corporate financial data, and a method using the optimal threshold to improve AUC performance even when the total imbalanced data are used.

Figure 6 depicts four ROC AUC curves for four small random samples where a small random sample is obtained by random sampling of the non-bankrupt companies five times. From Figure 6, it can be seen that the performance of diagnosing non-bankrupt companies is excellent for k-NN and RF, and it is difficult to compare the performances of LR and DT. Figure 7 plots the ROC AUC curve the small sample generated using approximate entropy, where RF had the best performance in diagnosing non-bankrupt companies, followed by k-NN, LR, and DT. These results are consistent with those in Panels (a) and (b) in Table 4. Figure 8 represents the optimal threshold identified based on the AUC for each model as a dot.

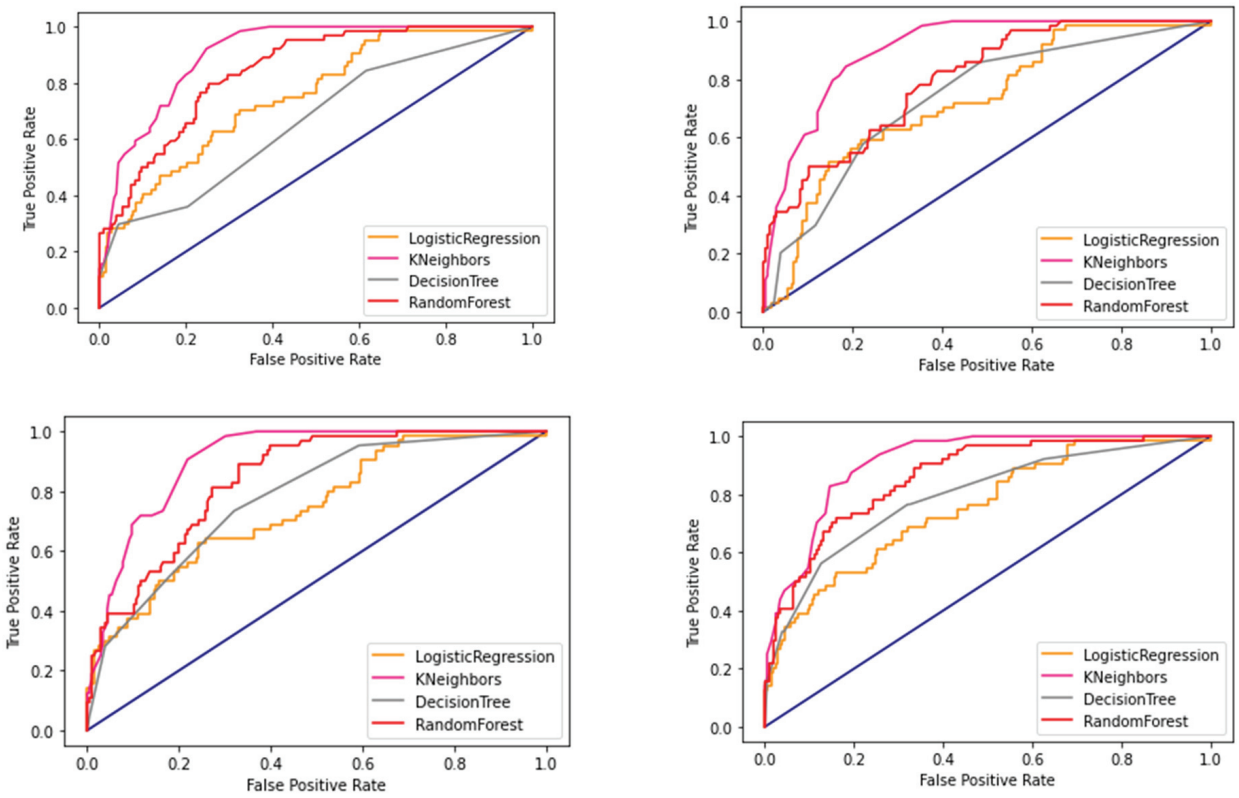


Figure 6. ROC AUC graph for four small data samples sampled randomly.

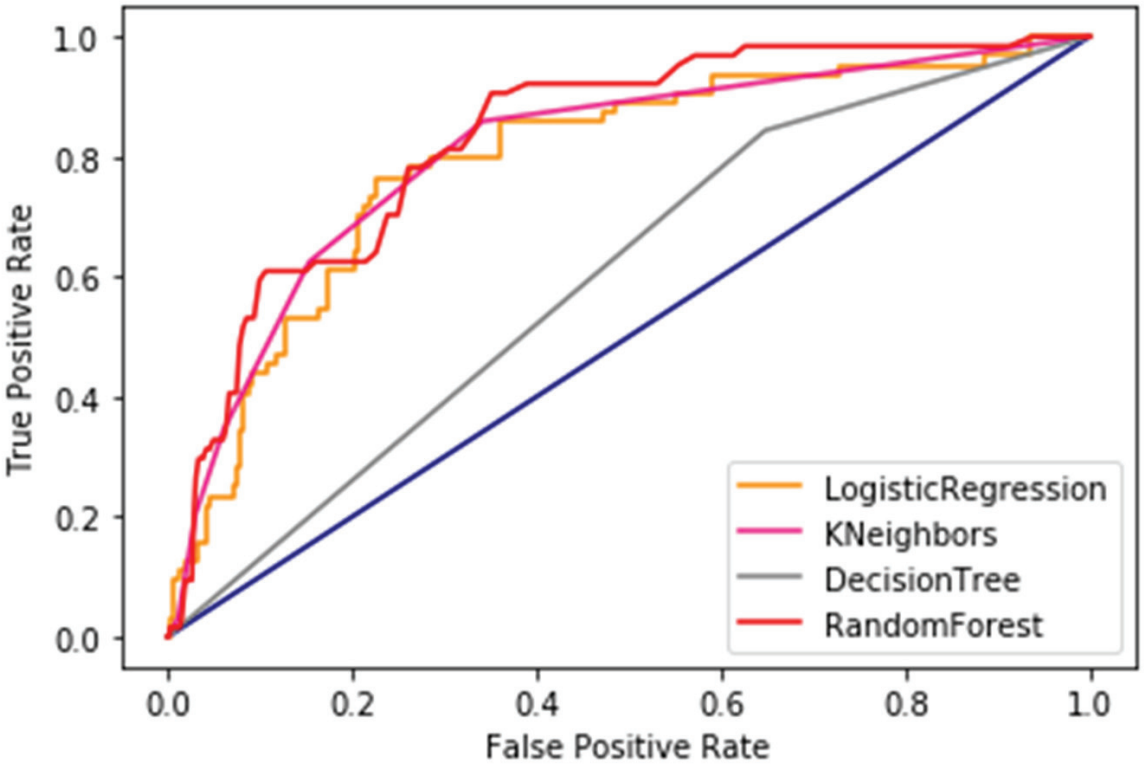


Figure 7. ROC AUC graph for the small sample created using approximate entropy.

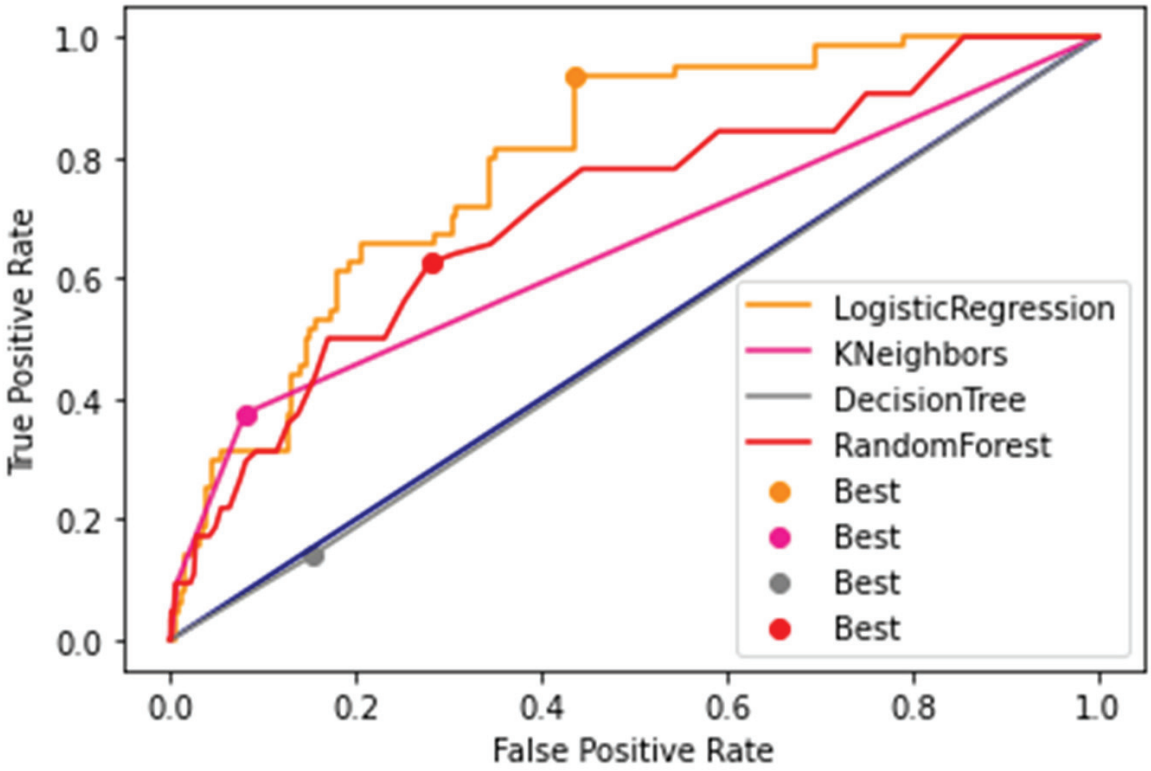


Figure 8. ROC AUC curve with the highest score (optimal threshold = 0.4357 for the LR, 0.0800 for k-NN, 0.1326 for DT, and 0.3789 for RF).

4. Conclusions

We collected 13 financial information items for about 1020 companies listed on KOSPI and KOSDAQ from 2012 to 2021 and compared the bankruptcy prediction performance of the LSTM, LR, k-NN, DT, and RF models.

As the collected financial data on the companies in Korea represents a small sample due to the small number of bankruptcies, the LSTM is not a suitable corporate bankruptcy prediction model.

As the LR, DT, RF, and k-NN classification models do not use time series data as an input, we configured financial information in year n and bankruptcy or non-bankruptcy in year $n+1$ as the input and label data, which made it possible to generate nine times more input data compared to the LSTM model and resulted in higher bankruptcy prediction performance.

The analysis of small sample data compared the random sampling method of non-bankrupt companies in the training dataset with the sampling method using approximate entropy, where the latter outperformed the former method in bankruptcy precision and bankruptcy recall. The performance measures of the models trained with the dataset constructed using five random samples and those generated using approximate entropy were higher than those of the models trained with the sample data identified from [7], confirming that our proposed method can improve the bankruptcy prediction performance using a small sample relative to a large sample for an imbalanced dataset. In particular, for the unbalanced financial data collected in this study, our analysis proved that the AUC performance measure can be greatly improved by classification using the optimal threshold identified based on the AUC, as shown in Panel (c) in Table 4. The results of this study provide useful information for selecting a suitable bankruptcy prediction model when using a dataset with relatively few bankrupt companies.

In this study, the number of bankrupt companies for which financial information was collected is very small, so we need to use a strategy such as the cross-validation technique to reduce the variation that may occur when performing experiments by sampling non-bankrupt companies. We plan to conduct experiments applying these techniques in future works. We measured the performance by averaging the experimental results of random sampling with five repeated experiments, and it is necessary to increase the number of such repeated experiments to improve the reliability of the experimental results, so we plan to do so in future works. In this study, we did not conduct an experiment to compare the experimental results of the benchmark model and benchmark dataset with the results of this study, so we plan to add them in future works to ensure the reliability of the research results.

Funding: This research received no external funding.

Data Availability Statement: The code and datasets used and analyzed during this study can be obtained from the online resource (https://github.com/shnoh92/Corporate-Bankruptcy-Prediction-Models-Based-on-Imbalanced-Financial-Data/ accessed on 5 March 2023).

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

Table A1. Twenty-nine variables that correspond to the criteria characterizing bankrupt and non-bankrupt companies.

Category	Section	Feature
Financial Statements	Balance Sheet (1000 won)	Accumulations
		Retained Earnings
		Net assets of controlling shareholders (before capital stock reduction)
		Owners of Parent Equity
		Total Equity
	Comprehensive Income Statement (1000 won)	Earnings before tax
		(Total Comprehensive Income Attributable to) Owners of Parent Equity
		Total Comprehensive Income
	Cash Flow Statement (1000 won)	Cash Flow
Financial Ratio	Stability (%)	Intangible Asset Ratio
		Equity Capital Ratio
		Borrowings and Bonds Payable Ratio
		Borrowed Capital Ratio
		Cash Flow/Total Debt
		Cash Flow/Total Equity
		Cash Flow/Total Asset
	Growth (yearly) (%)	Total Asset Growth Rate
	Profitability (%)	Operating Revenue/Operating Expense
		Profit Margin Ratio
		ROA (Current Net Income)
		ROA (Earnings before tax)
		ROA (Operating Profit)
		ROA (Total Comprehensive Income)
		ROE (Current Net Income)
		ROE (Earnings before tax)
		ROE (Operating Profit)
		ROE (Net profit of controlling shareholders)
	Activity (times)	Total Debt Turnover
		Total Asset Turnover

References

1. Oh, W.S.; Kim, J.H. Forecasting corporate bankruptcy with artificial intelligence. *J. Ind. Converg.* **2017**, *15*, 17–32.
2. Cha, S.; Kang, J. Corporate default prediction model using deep learning time series algorithm, RNN and LSTM. *J. Intell. Inf. Syst.* **2018**, *24*, 1–32.
3. Barboza, F.; Kimura, H.; Altman, E. Machine learning models and bankruptcy prediction. *Expert Syst. Appl.* **2017**, *83*, 405–417. [CrossRef]
4. Falavigna, G. Financial ratings with scarce information: A neural network approach. *Expert Syst. Appl.* **2012**, *39*, 1784–1792. [CrossRef]
5. Hinton, G.E.; Osindero, S.; The, Y.-W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef] [PubMed]
6. Jang, Y.; Jeong, I.; Cho, Y.; Ahn, H. Business Failure Prediction with LSTM RNN in the Construction Industry. In Proceedings of the ASCE 2019 International Conference on Computing in Civil Engineering, Atlanta, GA, USA, 17–19 June 2019; pp. 1–8.

7. Kim, H.; Cho, H.; Ryu, D. Corporate bankruptcy prediction using machine learning methodologies with a focus on sequential data. *Comput. Econ.* **2022**, *59*, 1231–1249. [CrossRef]
8. Odom, M.D.; Sharda, R. A neural network model for bankruptcy prediction. In Proceedings of the 1990 IJCNN International Joint Conference on Neural Networks, San Diego, CA, USA, 17–21 June 1990; pp. 163–168.
9. Wilson, R.L.; Sharda, R. Bankruptcy prediction using neural networks. *Decis. Support. Syst.* **1994**, *11*, 545–557. [CrossRef]
10. Kim, H.; Cho, H.; Ryu, D. Corporate default predictions using machine learning: Literature review. *Sustainability* **2021**, *12*, 6325. [CrossRef]
11. Brygata, M. Consumer Bankruptcy Prediction Using Balanced and Imbalanced Data. *Risks* **2022**, *10*, 24. [CrossRef]
12. Zhou, L. Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. *Knowl. Based Syst.* **2013**, *41*, 16–25. [CrossRef]
13. Garcia, V.; Jose, S.S.; Ramon, A.M. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowl. Based Syst.* **2012**, *25*, 13–21. [CrossRef]
14. Syed, N.; Sharifah, H.; Shafinar, I.; Bee, W.Y. Personal bankruptcy prediction using decision tree model. *J. Econ. Financ. Adm. Sci.* **2019**, *24*, 157–170. [CrossRef]
15. Amidon, A. PyOD: A Unified Python Library for Anomaly Detection. 11 May 2021. Available online: https://towardsdatascience.com/pyod-a-unified-python-library-for-anomaly-detection-3608ec1fe321 (accessed on 15 January 2023).
16. Mishra, S.; Kshisagar, V.; Dwivedula, R.; Hota, C. Attention-Based Bi-LSTM for Anomaly Detection on Time-Series Data. In Proceedings of the 2021 ICANN International Conference on Artificial Neural Networks, Online, 14–17 September 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 129–140.
17. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In Proceedings of the 2021 NeurIPS 35th Conference on Neural Information Processing Systems, online, 6–14 December 2021; pp. 1–14.
18. Richman, J.S.; Moorman, J.R. Physiogical time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circ. Physiol.* **2000**, *278*, H2039–H2049. [CrossRef] [PubMed]
19. Noh, S.-H. Analysis of gradient vanishing of RNNs and performance comparison. *Information* **2021**, *12*, 442. [CrossRef]
20. Jagannath, V. Random Forest Template Tibco Spotfirer Wiki Page. 24 March 2017. Available online: https://community.tibco.com/wiki/random-forest-template-tibco-spotfirer-wiki-page (accessed on 15 January 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

sustainability

MDPI

Article

Financial Technology Development and Green Total Factor Productivity

Wentao Hu [1,*] and Xiaoxiao Li [2]

[1] The School of Economics, Xiamen University, Xiamen 361005, China
[2] School of Mathematics and Statistics, Shangqiu Normal University, Shangqiu 476000, China; lixiaoxiao@sqnu.edu.cn
* Correspondence: 15620190154369@stu.xmu.edu.cn

Abstract: As a new product resulting from the deep integration of the financial industry and artificial intelligence (AI) technology, financial technology (fintech) has a significant impact on the progress of green total factor productivity (GTFP). Based on city-level data from 2011 to 2021 in China, this paper used the super-efficiency SBM model with embedded non-expected output and the GML index method to measure the GTFP levels of 283 prefecture-level and above cities and to empirically test the impact of fintech on GTFP and its underlying mechanisms. The empirical results showed that the development of fintech had significantly promoted the improvement of GTFP, and the effect was dynamically stable. Specifically, fintech had a stronger and more significant incentive effect on GTFP in its more mature stage of development. By decomposing fintech into two dimensions, it was found that the depth of fintech development had a stronger impact on GTFP with dynamic superimposed characteristics. Mechanism analysis showed that fintech development can drive the progress of GTFP by improving resource allocation efficiency, optimizing human capital, and incentivizing technological innovation channels. Moderating effect analysis revealed that financial regulation and environmental regulation have a positive moderating effect on the baseline relationship between fintech and GTFP. Further research found that the moderating effects of financial regulation and environmental regulation exhibit significant nonlinear threshold characteristics, and the driving effect of fintech on GTFP can only reach its maximum when both are within the optimal range. This study provides valuable insights for the development and optimization of fintech, the green transformation of the real economy, and high-quality development.

Keywords: financial technology; financial regulation; environmental regulation; green total factor productivity

1. Introduction

China, as a typical developing country, has gained tremendous economic achievements and significantly improved its comprehensive national strength during more than 40 years of reform and opening up. However, objectively speaking, the rapid growth of the past decades has been based on the extensive development model that relies on the consumption of a large amount of material resources and pursues quantity and spatial expansion [1]. This extensive growth path has overemphasized the speed of economic growth while relatively neglecting the quality of economic development, which has led to numerous resource, environmental, ecological, and social problems such as tightened resource constraints, severe environmental damage, and low economic efficiency [2]. In recent years, with the profound changes in the international market environment and domestic factor endowment conditions, the previous high-input, high-consumption, and high-pollution extensive development is unsustainable, and the economic development model urgently needs to transform "rough" development with an emphasis on factor

inputs into "high-quality" development that integrates environmental protection and economic efficiency improvement [3]. Green total factor productivity (GTFP), which takes into account the improvement of ecological and economic efficiency, coincides with the concept of high-quality economic development and has become an important tool to improve the quality of economic development [4]. In simple terms, GTFP refers to the efficiency level of an economic system in utilizing production factors while considering environmental factors. By incorporating environmental benefits into the calculation of total factor productivity, it reflects the performance of the economy in environmental protection and sustainable resource utilization [4]. Enhancing GTFP is of paramount importance in achieving high-quality economic development, promoting the transition to a green economy, and facilitating sustainable development in China. By improving resource efficiency and reducing environmental pressures, China can achieve a harmonious balance between economic growth and environmental protection, making positive contributions to sustainable development [5]. Recently, the Chinese government has attached great importance to the improvement of GTFP and the quality of economic development, repeatedly emphasizing the adherence to the new development concept of "innovation, coordination, green, openness, and sharing". The "14th Five-Year Plan for National Economic and Social Development and the Outline of the Long-Range Objectives for 2035" issued in March 2021 further stressed the need to "promote the green transformation of economic and social development and build a beautiful China", and made high-quality economic development the theme and key goal of economic and social development in the 14th Five-Year Plan and even in the longer term. GTFP can be considered the concentrated summary of high-quality economic development [6], which profoundly reveals the new direction and new momentum of economic development, and is of great practical significance to China in terms of accelerating the transformation of the economic development mode, as well as promoting the upgrading of economic quality, power, and even efficiency [7].

As the lifeblood of the real economy, finance is an important condition to promote innovative development and efficiency of the real economy, and an incomplete financial market is a key factor hindering the improvement of total factor productivity [8], while the integration of artificial intelligence (AI) technology with financial support can help improve the quality of financial service supply and resource allocation efficiency [9]. Generally speaking, AI is a comprehensive technological field that encompasses the integration of various technologies such as big data, cloud computing, blockchain, etc. It aims to develop intelligent algorithms and techniques that enable computer systems to perceive, understand, learn, and make decisions. In recent years, with the widespread application of AI technology, China's financial technology (fintech) has developed by leaps and bounds. According to statistics, by the end of 2021, more than 20,000 companies had applied for fintech-related patents in China, with the total number of patents reaching more than 120,000, and the number and growth rate of patents in the past five years were far higher than those of other countries [10]. From the perspective of business, China's fintech is more focused on technology "empowerment", that is, leveraging the deep integration of AI technology and financial services to upgrade financial products and business processes, solving long-standing problems in the financial market such as information asymmetry and high operating costs, and thereby reducing financial service thresholds, expanding the accessibility of financial services, and providing intrinsic motivation for improving the quality and efficiency of the real economy [9]. Can the financial industry innovation driven by the development of AI technology boost GTFP growth? In the context of economic structural transformation and upgrading, deepening financial reform, and arduous ecological and environmental protection tasks, the exploration of this issue undoubtedly has important theoretical and practical values.

At present, the academic community pays much attention to topics related to fintech or digital finance. Numerous studies have analyzed the impact of fintech on total factor productivity (TFP) from micro or macro perspectives. From a micro perspective, fintech can improve enterprise TFP through various channels, such as alleviating information

asymmetry between banks and enterprises and easing financing constraints [11], improving credit resource allocation efficiency [10], enhancing enterprise revenue and financial efficiency [12], and increasing enterprise innovation capabilities [13]. From a macro perspective, the application of financial technologies such as big data and cloud computing is conducive to risk identification and management, reducing risk concentration, ensuring the stability of the financial system, and thus improving resource allocation efficiency [14]. The study by Hou Cen and Li Beiwei (2020) [15] showed that the role of fintech in improving TFP is mainly achieved through mechanisms such as enhancing technology spillover and promoting industrial structure upgrading. Additionally, Wang Xin (2015) [16] argued that internet-based financial services help to alleviate credit allocation and promote rational credit resource allocation, driving the progress of TFP. Bunea et al. (2016) [17] pointed out that fintech innovation poses certain challenges to the dominant position of traditional financial institutions, restructuring the traditional resource allocation model based on indirect financing, but traditional financial institutions can also use financial technology to optimize financial products and services, achieving a new growth curve. In recent years, with the rapid integration of digital technology and the real economy, some scholars no longer stick to the specialized category of fintech and have started to examine the influencing factors of TFP from a broader digital perspective, whose mechanisms of action can be broadly categorized as technological progress, labor efficiency improvement, and industrial agglomeration [18,19]. Throughout the existing studies, only a small amount of studies have investigated the impact of fintech or digital finance on GTFP, and it mainly focused on single causal identification, spatial spillover transmission, and regional heterogeneity [20–22], and the discussion of the underlying mechanisms is still insufficient. Therefore, it is necessary to further explore the "black box" of the mechanisms through which fintech affects GTFP and scientifically evaluate its role through empirical tests, which can help deepen the academic understanding of the economic and environmental impacts of fintech.

It is worth noting that the government's rational understanding of the ecological environment is also essential to achieve the harmonious development of the social economy and ecological environment and thus to promote the continuous progress of GTFP. In a market economy, due to the negative externalities of environmental pollution, it is difficult for economic actors to adopt systematic environmental protection strategies [23]. Therefore, environmental regulation at the government level becomes an important means to promote environmental governance and high-quality economic development [24]. Theoretically speaking, as a mandatory measure, environmental regulation can promote the development of clean environmental industries and force the green transformation of polluting industries, thus further driving GTFP [5]. In addition to actively embracing frontier technologies to help upgrade industries and transform the economy, another important reason for China's tremendous achievements in green industry development and ecological environment management in recent years is the Chinese government's high regard for environmental protection and increasingly tightening environmental regulation policies [20]. Moreover, China is a vast country, and the intensity of environmental regulations naturally varies greatly across regions. Therefore, it is necessary to ask what role environmental regulation plays in the relationship between fintech and GTFP, and whether the impact of fintech on GTFP is heterogeneous depending on the intensity of environmental regulation. The current literature does not provide a systematic response to this question. At the same time, fintech does not change its own financial logic, let alone the basic core of "risk-return" in the financial industry. Fintech innovation will also generate new risks, and with AI technology, the financial risks embedded in fintech will be transmitted more quickly and through multiple channels, which will eventually have a dampening effect on the productivity of the real sector [13]. Therefore, fintech undoubtedly has the dual attributes of both efficiency gain and risk induction. Due to the existence of this dual attribute, effective regulation is essential [9]. Financial regulation facilitates the sustainable development of fintech. Regulatory bodies establish and enforce laws and policies that require fintech companies to adhere to environmental protection and sustainable development principles

while promoting GTFP. These regulations may include disclosure requirements for fintech firms to report their environmental and social impacts, ensuring that their operations and innovations do not have adverse effects on the environment [25]. In this regard, it is worth exploring the role of financial regulation in the "fintech–GTFP" relationship. This will help achieve benign development of fintech, improvement of financial regulatory models, and enhancement of the quality and efficiency of the real economy. Table 1 presents some prior studies on this topic.

Table 1. Overview of the relevant studies related to the research topic of this paper.

Author(s)	Fintech/Digital Finance	Financial Regulation	Environmental Regulation	Mechanism Effect	GTFP
[9]	Yes	No	No	Yes	No
[13]	Yes	Yes	No	Yes	No
[10]	Yes	No	No	Yes	No
[12]	Yes	No	No	Yes	No
[5]	No	No	Yes	Yes	Yes
[4]	No	No	No	No	Yes
[6]	No	No	Yes	Yes	Yes
[17]	Yes	Yes	No	Yes	No
[21]	Yes	No	Yes	No	Yes
[26]	No	No	No	No	Yes
[27]	Yes	No	No	Yes	No
[24]	Yes	No	Yes	Yes	No
[28]	Yes	No	No	Yes	No
[29]	No	No	Yes	Yes	No
[30]	Yes	No	No	Yes	No
[31]	Yes	Yes	No	Yes	No
[22]	Yes	No	No	Yes	Yes
[32]	Yes	Yes	No	Yes	No
[14]	Yes	Yes	No	No	No
This study	Yes	Yes	Yes	Yes	Yes

Compared with the existing literature, the marginal contributions of this paper are mainly in the following aspects: First, in terms of the research objective, the existing literature mainly discusses the impact of fintech or digital finance on TFP, while this paper focuses on the relationship between fintech and GTFP, and for the first time examined the dynamic effects and nonlinear effects of fintech on GTFP, thus creating an important supplement to the literature on financial functions and providing new empirical evidence for clarifying the economic effects of fintech. Second, under a unified framework, this paper explored the heterogeneity of fintech's impact on GTFP, comprehensively and systematically evaluating the linear and non-linear regulatory roles played by environmental regulation and financial supervision. This provides important empirical evidence for China in formulating refined policies to support the progress of GTFP and the upgrading of economic development quality. Third, we analyzed the mechanisms and paths of the impact of fintech on GTFP from three aspects: resource allocation, technological innovation, and human capital, further opening up the "black box" of the inherent mechanisms of financial technology, and providing new evidence for a deeper understanding of the impact of emerging financial forms on green economic development.

2. Theoretical Analysis and Research Hypothesis

Based on the existing research, this study investigated and demonstrated the impact of fintech on GTFP from two aspects: the mechanism of action and the regulatory effect. Based on this, the research hypotheses of this article were proposed. Before theoretical analysis, Figure 1 presents the basic conceptual framework and research hypotheses of this study.

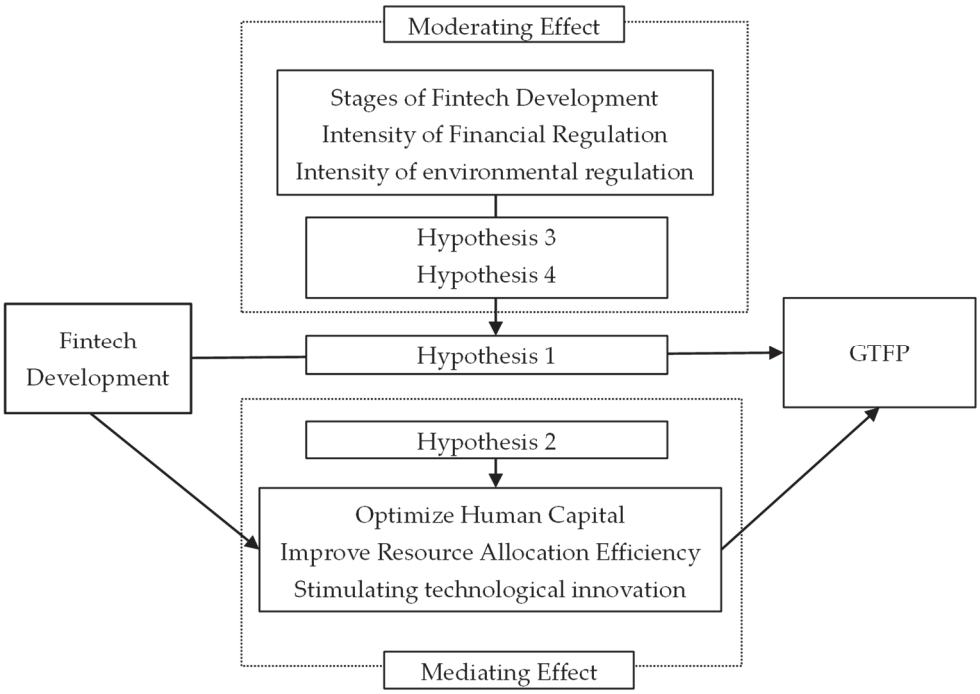


Figure 1. The conceptual framework of fintech on GTFP. Note: The framework diagram includes this study's main research questions and hypotheses.

2.1. The Mechanism of Fintech's Effects on GTFP

Green total factor productivity (GTFP), which takes environmental factors into account, is an important indicator of the quality of economic development in a country or region [4]. As resource and environmental constraints become increasingly tight and ecological carrying capacity approaches saturation, a country or region can only achieve green and sustainable development by transforming to endogenous growth that relies on GTFP [5]. Among the many factors affecting GTFP, the completeness of financial markets and their effectiveness in resource allocation are undoubtedly the key variables. A well-functioning financial market can not only provide sufficient and timely financial support for factor inputs and technological R&D of market players but also allocate capital to key segments and critical sectors of the economy, thereby improving the level of social production [12]. China's financial structure is dominated by bank-led indirect finance, coupled with the tendency toward financial system and government credit regulation, which has led to long-standing issues such as information asymmetry, structural mismatch, and distortion in resource allocation, becoming important factors hindering the improvement of GTFP [10]. However, in recent years, seizing the great opportunities brought by the rapid development of AI technologies such as big data and cloud computing, the finance sector has also strengthened its integration with AI technology. Fintech has emerged and developed rapidly, and the challenges faced by traditional financial development are expected to be reversed with the "empowerment" of fintech [9].

Fintech is the product of the deep integration of finance and AI technology. Relying on underlying AI technologies such as big data, cloud computing, and blockchain, it can effectively alleviate information asymmetry, which helps to improve resource allocation efficiency and technological innovation capabilities [31], thereby providing endogenous power for the upgrading of GTFP. Technological innovation is an important factor affecting GTFP, while financing constraints and credit allocation distortions can severely constrain technological R&D [33]. The synergistic development of fintech and traditional financial institutions can significantly increase market transparency and curb credit allocation imbalance [32], which can be beneficial in alleviating the financing constraints faced by enterprises [34]. Fintech in China has both technology spillover effects and inclusive effects,

not only significantly improving the total factor productivity of commercial banks [35], but also having a stronger promotion effect for innovation in areas with less developed financial sectors [36]. The combination of AI technology and finance can also use both "incremental supplementation" and "stock optimization" to break down technological innovation barriers [13]. On the one hand, fintech can efficiently integrate financial resources, allocate them rationally to the demand side, and enrich enterprise financing methods and identify the optimal R&D paths for enterprises by providing high-quality data analysis tools and intelligent products, thereby driving the creation of new knowledge and technological iterations [37]. On the other hand, fintech can effectively drive the upgrading of traditional financial structures, improve the operation modes and pricing models of traditional financial products, and provide a favorable financial environment for innovation activities [27].

Resource allocation efficiency is also a key factor affecting GTFP. Hsieh and Klenow (2009) [38] argued that resource allocation distortion is a major constraint to factor productivity growth in China. If China's resource allocation efficiency could reach the level of the United States, its productivity would increase by 30–50%. Numerous studies have shown that fintech has significant advantages in reducing resource allocation distortions and improving resource allocation efficiency. Firstly, fintech can not only solve the trust issue between transaction parties based on the tamper-proof characteristics of blockchain, promoting the smooth completion of transactions and the rational flow of production factors [39], but also overcome information asymmetry in the transaction process through big data analysis and information processing technology, effectively matching supply and demand of funds, expanding the boundaries of financial services, and improving resource allocation efficiency [30]. Secondly, fintech, characterized by high synergy and networking, can accelerate information interconnectivity between institutions and regions, expand resource allocation boundaries, and promote the free flow of talents, capital, and technology across industries, sectors, and regions [40]. Thirdly, the innovative and integrated development of finance and AI technology can enhance the vitality and efficiency of the financial market, improve the traditional financial sector's ability to consolidate, organize, and analyze information, help the real sector overcome transaction frictions, information asymmetry and other problems, and improve the effectiveness of investment and financing decisions in the real economy [13]. Finally, fintech significantly reduces information asymmetry costs in long-tail markets by collecting, analyzing, and sharing real-time and intelligent basic information of financial and non-financial institutions, using decentralized distributed ledger technology to ensure information security and reduce trust risks, thereby expanding the scope of financial services to small and medium enterprises (SMEs). It enhances the identification of outstanding enterprises with development potential but faces budget constraints, and ultimately strengthens the allocation efficiency of financial resources for technological innovation projects [41].

Among the many factors that determine GTFP, the important role played by the accumulation level and quality of human capital cannot be overlooked [42]. Human capital condensed with educational inputs is the basis of technological innovation and naturally facilitates the development and output of green technologies [43]. At the same time, higher levels of human capital generally have stronger technology absorption capabilities, enabling better application of green technologies in economic practice [44]. Moreover, human capital affects the consumption of green products in society, and a high level of human capital tends to have a stronger tendency to consume innovative products [45], which is extremely beneficial in driving the overall progress of GTFP. Meanwhile, fintech can also drive the accumulation level and quality of human capital from multiple perspectives. First, the development of fintech has greatly expanded the customer base that can be reached by financial services, effectively reducing the financing constraint of residents and improving the availability of credit, which is conducive to stimulating the consumption propensity of residents in knowledge acquisition, education, and training [9]. Second, fintech can also provide more investment channels for low-income groups excluded by traditional finance,

realize individual wealth appreciation, narrow the income gap between regions [12], and lay a solid material foundation for human capital accumulation. Third, the information technologies such as big data and cloud computing used in fintech rely on the Internet, and the Internet has a significant network effect. Fintech innovation activities can break the temporal and spatial boundaries of knowledge dissemination and accelerate knowledge exchange and communication through network effects, and market subjects can imitate advanced technologies, learn advanced knowledge, and generate new knowledge, which greatly contributes to the optimization of the human capital structure [46]. As a result, this paper proposes:

Hypothesis 1. *Fintech development contributes to GTFP enhancement.*

Hypothesis 2. *Fintech development can contribute to GTFP progress by stimulating technological innovation, improving resource allocation efficiency, and optimizing human capital.*

2.2. Nonlinear Relationship between Fintech and GTFP

As analyzed earlier, fintech has the advantages of high efficiency, low cost, and wide service range, which can promote GTFP through encouraging technological innovation, improving resource allocation efficiency, and optimizing human capital. However, the development of fintech is a gradual process with the characteristic of "one hair moves the whole body", and its impact on GTFP may be characterized by phases [21]. On the one hand, in the early stage of fintech development, due to the lack of synchronization and automation of financial regulation, the rapidly updated fintech products may have a high risk of uncertainty [19], which poses a negative impact on the real sector. With the gradual maturity of fintech and the improvement of the financial regulatory system, the service function of fintech can be gradually optimized [27], and its contribution to GTFP will be increasingly enhanced. On the other hand, with the widespread application of fintech, the marginal cost of linkage among various sectors continues to decline, while the benefits obtained by participants increase geometrically, and this effect will become more and more obvious as the level of fintech increases [47], that is, the "Metcalfe's Law" and "Moore's Law" in the internet field hold in the relationship between fintech and GTFP [27]. Therefore, this article proposes:

Hypothesis 3. *The impact of fintech on GTFP exhibits heterogeneity in different stages of its own development.*

2.3. The Moderating Effect of Fintech on GTFP

As analyzed earlier, fintech does not change the inherent logic of finance itself, nor does it change the basic rule of "risk-return" in the financial industry. Fintech is essentially a product of the deep integration of AI technology and traditional finance, and innovation is the inherent driving force of fintech, which constitutes an important component of financial risk while promoting financial development [9]. Fintech innovation will increase the probability of financial risk contagion, which may expand the risk exposure of the real sector and adversely affect GTFP. As a result, fintech has both efficiency-enhancing and risk-inducing properties, so it is a necessary task to impose effective regulation on it [27]. This matters for the effectiveness of fintech, the stability of financial markets, and the quality improvement of macroeconomic development. In fact, any financial innovation, including fintech, cannot avoid the theoretical detour of financial innovation, financial risk, and financial regulation [10]. On the one hand, strengthening financial regulation can undoubtedly help guard against systemic financial risks and enhance the positive effects of fintech; on the other hand, the uncertainty and complexity of fintech may also amplify the shortcomings of financial regulation, and inappropriate financial regulation may also fetter the release of fintech effectiveness and impede the improvement of GTFP.

To promote GTFP progress and thus achieve sustainable development in harmony with nature, the pushback effect of environmental regulation is indispensable [48]. The non-exclusivity and non-competitiveness of the environment and the negative externalities of pollution lead to market failure in environmental protection, and government environmental regulation policies are crucial to internalize environmental costs [5]. The rapid development of the internet and digital technologies has greatly expanded channels for environmental governance and social participation, effectively enhancing the effectiveness of environmental regulation [49]. Environmental regulation policies also produce an indicative effect. Under the guidance of environmental regulation, financial resources mobilized by fintech will be invested more in green and environmentally friendly, clean energy, and resource-intensive industries, which is conducive to achieving a win–win situation among ecological, economic, and social benefits and providing strong support for sustainable high-quality development [24]. However, there is no consensus among academics on the relationship between environmental regulations and GTFP. When environmental regulation reflects resource scarcity, it generates a so-called "compliance cost" where increased spending on environmental protection may crowd out production resources and reduce innovation returns, thus having a negative impact on GTFP [50]. However, the "Porter Hypothesis" argues that, under appropriate constraints on environmental regulation, firms will increase green technology development and reduce environmental pollution in the production process, partially or fully offsetting the loss of environmental regulations [51], which will significantly increase GTFP. Therefore, the impact of environmental regulation on GTFP may not be a simple linear relationship. Based on the above analysis, this article proposes:

Hypothesis 4. *Financial regulation and environmental regulation play a moderating role between fintech development and GTFP.*

3. Econometric Model, Variables, and Data

3.1. Construction of the Econometric Model

Based on the theoretical analysis and existing literature discussed earlier, this paper constructed the following baseline econometric model to investigate how the development of regional fintech affects the level of a city's GTFP:

$$GTFP_{i,t} = \alpha_0 + \alpha_1 Fintech_{m,t-1} + \sum_{k=2}^{T} \alpha_k Controls_{i,t} + \mu_i + \delta_t + \varepsilon_{i,t} \quad (1)$$

where the dependent variable is the GTFP of city i in year t, which is measured based on the super-efficiency SBM model (slack-based measure) with GML (global Malmquist–Luenberger) index method for non-expected output. The core explanatory variable $Fintech_{m,t-1}$ denotes the level of fintech development of province m in year $t-1$, using the fintech development index as its proxy variable. To control for the influence of omitted variables, multiple city-level control variables, $Controls_{i,t}$ are included in this paper; the fixed effects for city and year are denoted by μ_i and δ_t, while $\varepsilon_{i,t}$ represents the random error term. To improve the estimation efficiency, cluster robust standard errors are used by default in the regression process. Additionally, considering that there is lag time between regional fintech development and its effects on urban GTFP, we followed an approach in a previous study [12] and used a lagged fintech development index, which can also alleviate the reverse causality problem. We focused on the coefficient of the core explanatory variable $Fintech_{m,t-1}$, and if α_1 is significantly positive, this indicates that the development of fintech can significantly enhance the level of urban GTFP, and Hypothesis 1 is supported.

3.2. Variable Definitions

3.2.1. Explained Variable

GTFP incorporates energy consumption and environmental pollution into the economic growth framework, emphasizing the green concept of coordinated development of economy, resources, and environment, which is a significant improvement to traditional total factor productivity. The academic community's measurement of GTFP mainly includes methods such as the traditional radial DEA model [52], the SBM model with undesirable outputs [26], and the stochastic frontier analysis model [53]. However, these methods have problems, mainly, ignoring slack variables, effective DMU is not easy to distinguish, and poor comparability across periods. Therefore, based on the research of Tone (2001) [26] and Oh (2010) [54], this paper uses the super-efficiency SBM model that considers energy consumption and environmental pollution, and applies the GML index of global parameter ratio to measure the GTFP of 283 cities in China, denoted as GTFP_GML. This method can overcome the three key issues of variable relaxation, efficient DMU distinguishability, and cross-period comparability, ensuring the robustness of the results. In addition, the SBM-GML approach to measure GTFP requires determining input and output indicators; for these, this paper referred to Fare et al. (2009) [55], taking labor, energy, and capital as input indicators, real GDP as the expected output indicator, and industrial SO_2 emissions, industrial wastewater emissions, and industrial smoke and dust emissions as undesirable outputs. The relevant indicators and metrics are shown in Table 2. The specific measurement method can be divided into the following three steps:

Step 1: Define the environmental technology set for measuring GTFP:

$$\mathrm{ETS}=\left\{\left(X,\ \overline{Y}^{g},\ \overline{Y}^{b}\right)\middle|\ \overline{X}\geq\sum_{\substack{j=1\\ j\neq 0}}^{L}\lambda_j x_j,\ \overline{Y}^{g}\leq\sum_{\substack{j=1\\ j\neq 0}}^{L}\lambda_j y_j^{g},\ \overline{Y}^{b}\leq\sum_{\substack{j=1\\ j\neq 0}}^{L}\lambda_j y_j^{b},L\leq e\lambda\leq\mu,\lambda_j\geq 0\right\} \tag{2}$$

where the input–output pattern of each city corresponds to n inputs, s_1 is the desired outputs, and s_2 is the non-desired outputs. X represents the n-dimensional input vector, $X=(x_1,x_2,\cdots,x_n)\in R_+^n$; Y^g represents the s_1-dimensional desired output vector, $Y^g=\left(y_1^g,y_2^g,\cdots,y_{s_1}^g\right)\in R_+^{s_1}$; Y^b represents the s_2-dimensional non-desired output vector, $Y^b=\left(y_1^b,y_2^b,\cdots,y_{s_1}^b\right)\in R_+^{s_2}$; and $\lambda=(\lambda_1,\lambda_2,\cdots,\lambda_L)$ represents the L-dimensional weight vector.

Step 2: Denote the excess input vector as $s^-\in R_+^{s_n}$, denote the excess of the undesired output as $s^b\in R_+^{s_2}$, and denote $s^g\in R_+^{s_1}$ as the deficiency of the desired output. A super-efficiency SBM model with undesired outputs is introduced based on the super-efficient SBM model:

$$\begin{aligned}\theta=\min_{\lambda,\,\overline{x},\,\overline{y}^g,\,\overline{y}^b}\ &\frac{\sum_{i=1}^{m}\frac{\overline{x}_t}{x_{io}}}{\frac{1}{s_1+s_2}\left(\sum_{r=1}^{s_1}\frac{\overline{y}_r^g}{y_{ro}^g}+\sum_{k=1}^{s_2}\frac{\overline{y}_k^b}{y_{ko}^b}\right)}\\ \text{s.t. } X\geq\sum_{\substack{j=1\\ j=0}}^{L}\lambda_j x_j,\ &\overline{Y}^g\leq\sum_{\substack{j=1\\ j=0}}^{L}\lambda_j y_j^g,\ \overline{Y}^b\geq\sum_{\substack{j=1\\ j=0}}^{L}\lambda_j y_j^b\\ &\overline{X}\geq x_o,\ \overline{Y}^g\leq y_o^g,\ \overline{Y}^b\geq y_o^b\\ &\overline{Y}^g\geq 0,\ \overline{Y}^b\geq 0,\ L\leq e\lambda\leq\mu,\ \lambda_j\geq 0\\ &\overline{x}_t=x_{io}+s^-(i=1,\cdots,n),\ \overline{y}_r^g=y_{ro}^g-s^g(r=1,\cdots,s_1),\ \overline{y}_k^b=y_{ko}^b-s^b(r=1,\cdots,s_1)\end{aligned} \tag{3}$$

where $\overline{x}$, $\overline{y}_r^g$, and $\overline{y}_k^b$ are the target values of the input and output of the evaluated unit, and x_{io}, y_{ro}^g, and y_{ko}^b represent the corresponding original values.

Step 3: According to the non-parametric framework and based on the above super-efficiency SBM model containing the undesirable outputs, construct a non-angular and non-radial MPI:

$$GML_o^{t,t+1} = \frac{\theta_o^{t+1}\left(x_o^{t+1}, y_o^{g,t+1}, y_o^{b,t+1}\right)}{\theta_o^{t}\left(x_o^{t}, y_o^{g,t}, y_o^{b,t}\right)} \times \left[\frac{\theta_o^{g}\left(x_o^{t+1}, y_o^{g,t+1}, y_o^{b,t+1}\right)}{\theta_o^{t+1}\left(x_o^{t+1}, y_o^{g,t+1}, y_o^{b,t}\right)} \times \frac{\theta_o^{t}\left(x_o^{t}, y_o^{g,t}, y_o^{b,t}\right)}{\theta_o^{g}\left(x_o^{t}, y_o^{g,t}, y_o^{b,t}\right)}\right] \quad (4)$$

where $\theta_o^{t}\left(x_o^{t}, y_o^{g,t}, y_o^{b,t}\right)$ and $\theta_o^{t+1}\left(x_o^{t+1}, y_o^{g,t+1}, y_o^{b,t+1}\right)$ represent the city efficiency values at periods t and t+1, respectively; $\theta_o^{g}\left(x_o^{t}, y_o^{g,t}, y_o^{b,t}\right)$ is the efficiency value based on global production technology and t-period input–output values; $\theta_o^{g}\left(x_o^{t+1}, y_o^{g,t+1}, y_o^{b,t+1}\right)$ is the efficiency value based on global production technology and the input–output values of period t+1. $\frac{\theta_o^{g}\left(x_o^{t+1}, y_o^{g,t+1}, y_o^{b,t+1}\right)}{\theta_o^{t+1}\left(x_o^{t+1}, y_o^{g,t+1}, y_o^{b,t}\right)}$ reflects the proximity of the frontier in period t+1 to the global frontier, and $\frac{\theta_o^{t}\left(x_o^{t}, y_o^{g,t}, y_o^{b,t}\right)}{\theta_o^{g}\left(x_o^{t}, y_o^{g,t}, y_o^{b,t}\right)}$ reflects the proximity of frontier period t to the global frontier. $GML_o^{t,t+1}$ is the change in city GTFP from period t to t + 1. $GML_o^{t,t+1} = 1$ indicates no change in GTFP; $GML_o^{t,t+1} > 1$ indicates an increase in GTFP; and $GML_o^{t,t+1} < 1$ indicates a decrease in GTFP.

In addition, considering that the GTFP indicator used in this paper is more concerned with the reduction of the unexpected output and drawing on the work of Fare et al. (2009) [55], we further estimated the city GTFP using the directional distance function of SBM that includes the unexpected output (SBM-DDF) as a robustness test indicator, denoted as GTFP_DDF.

Table 2. Description of input and output indicators.

Input/Output	Input/Output	Indicator Description
Input	Capital input	Referring to Shan Haojie (2008) [56], taking 2005 as the base period and using the perpetual inventory method calculate the capital input, the depreciation rate was set at 10.96%/billion CNY
	Labor input	Sum of the number of unit employees, private employees, and self-employed employees in each city/10,000 people
	Energy input	Electricity consumption of each city as a proxy variable for energy consumption /10,000 TCE
Expected output	Real GDP	Real GDP of each city (based on 2005) (billion CNY)
Non-desired output	Industrial sulfur dioxide emissions	Industrial sulfur dioxide (SO_2) emissions (million tons)
	Industrial wastewater emissions	Industrial wastewater emissions (million tons)
	Industrial soot emissions	Industrial soot emissions (million tons)

3.2.2. Core Explanatory Variable

The Global Financial Stability Board (FSB) defines fintech as a technological form of integrating finance and technology into financial services, capable of effectively reducing the operational costs of the financial system and enhancing the output and service efficiency of traditional finance through emerging AI technologies such as cloud computing, big data, blockchain. According to this definition, fintech is essentially a business model, technology application, and even processes and products based on technological means to promote financial innovation. The Digital Inclusive Finance Index, jointly compiled by the Digital Finance Research Center of Peking University and Ant Financial Services Group, provides a comprehensive measurement of China's fintech business penetration, service efficiency, and depth of technology application in three dimensions: the breadth of fintech coverage,

the depth of fintech use, and the degree of fintech services; thus, this index can better reflect the stage characteristics and evolution trends of China's fintech development in recent years and has been widely adopted by scholars [57,58]. Based on the above considerations, this paper selected the Digital Inclusive Finance Index (provincial level), jointly compiled by the Digital Finance Research Center of Peking University and Ant Financial Services Group in China, as a proxy variable for fintech development, denoted as Fintech_P. At the same time, in the robustness test, we also used the city-level digital inclusive finance index, denoted as Fintech_C, as a proxy variable for fintech development. To improve the accuracy and rationality of the findings, the above data were treated as logarithms.

3.2.3. Mediating and Moderating Variables

The mediating variables mainly included technological innovation, resource allocation, and human capital level. In terms of technological innovation, previous studies often used R&D investment as a proxy variable for technological innovation, but real-world technological innovation activities often have typical high-risk characteristics, and the effective transformation of R&D investment into innovation output often has greater uncertainty, which may overestimate technological innovation capabilities. In view of this, this paper selected the logarithm of the number of patents granted in each city as a measure of technological innovation. Among them, the logarithm of the total number of invention patents granted was used to measure the quality of technological innovation, which is denoted as Innova_Q; the logarithm of the total number of utility model patents and design patents granted was used to measure the quantity of technological innovation, denoted as Innoa_N. Resource allocation efficiency, following the research of Chen Yongwei and Hu Weiming (2011) [59], was measured using the absolute value of the resource mismatch index, with Capital_M representing the degree of material resource mismatch and Labor_M representing the degree of labor resource mismatch, and the larger the absolute value of the index, the more serious the degree of resource mismatch and the lower the efficiency of resource allocation. For the human capital level, based on existing research, we used the proportion of the total population with a college degree or above in each city to measure the degree of human capital accumulation, denoted as Human_E. Based on the number of students enrolled in each level of education (general higher education, general undergraduate, general secondary school, secondary vocational and technical, etc.) and referring to the method of Liu Zhiyong et al. (2018) [60], the degree of human capital structure upgrading in each city was calculated using the vector index method and denoted as Human_A.

The moderating variables included environmental regulation and financial supervision. For environmental regulation, considering data availability and following the approach of He Lingyun and Qi Xiaofeng (2022) [5], the four basic indicators of industrial wastewater discharge compliance rate, industrial sulfur dioxide removal rate, industrial soot removal rate, and comprehensive utilization rate of solid waste were selected. By standardizing and removing dimensions, the entropy method was used to calculate the weight coefficients of each basic indicator, and then the weighted sum was used to obtain the provincial environmental regulation intensity indicator, denoted as Eregulat. For financial regulation, the ratio of financial regulatory expenditure to financial industry value added in each province was taken as a proxy indicator, denoted as Fsupervis, following Tang Song et al. (2020) [13].

3.2.4. Control Variables

In order to minimize the estimation bias caused by omitted variables, based on existing research, this study incorporated multiple city-level variables into the baseline model. These included: (1) government size (Gover), measured by the proportion of government fiscal expenditure to GDP; (2) trade openness (Ecopen), measured as the ratio of total imports and exports to GDP; (3) urbanization level (Urban), calculated as the ratio of urban non-farm population to total population; (4) industrialization level (Indus), measured as

the ratio of industrial added value to GDP; (5) financial development (Fincial), expressed as the ratio of total institutional deposits and loans to the city GDP; (6) industrial structure (Indstr), measured as the proportion of secondary industry output to GDP; (7) population density (Popul), expressed as the number of people (in hundreds) per square kilometer of the city; (8) greening level (Green), represented as the greening coverage rate in built-up areas; (9) foreign direct investment (Fdi), represented as the ratio of total foreign direct investment to local GDP; and (10) infrastructure (InfStra), using the road construction area per capita as a proxy variable.

3.3. Sample Selection and Data Sources

This paper used panel data from 283 prefecture-level and above cities in China from 2011 to 2021 to obtain a final sample of 3113 "city-year" observations. The fintech development index (provincial and city levels) was derived from the "Digital Inclusive Finance Index" from Peking University in China. Provincial-level data were mainly obtained from the "China Statistical Yearbook", "China Financial Statistical Yearbook", and "China Energy Statistical Yearbook"; city-level data were mainly taken from the "China City Statistical Yearbook", "China Environmental Statistical Yearbook", "China Energy Statistical Yearbook", "China Urban Construction Statistical Yearbook", and the CNRDS database. Table 3 gives the descriptive statistics of the main empirical variables in this paper. The mean value of GTFP calculated using the SBM-GML model was 1.0050 and the median value was 1.0011, indicating that the results were not significantly biased. The mean value of the fintech development index was 5.1464, which is slightly lower than its median of 5.3532, indicating a certain left skewness. The statistical information of the other control variables is consistent with previous research.

Table 3. Descriptive statistics of variables.

Variable	N	Mean	Sd	Min	P50	Max
GTFP_GML	3113	1.0050	0.0520	0.4884	1.0011	1.6335
Fintech_P	3113	5.1464	0.6595	2.9085	5.3532	6.0168
Fintech_B	3113	4.9872	0.7903	0.6729	5.2406	5.9523
Fintech_D	3113	5.1414	0.6139	1.9110	5.2459	6.0866
Innova_Q	3113	4.8298	1.7966	0.0000	4.5850	10.8770
Innova_N	3113	7.0968	1.5860	2.6391	6.9745	11.4340
Capital_M	3113	0.3764	0.3221	0.0002	0.3252	3.3634
Labor_M	3113	0.3682	0.3100	0.0007	0.2988	3.1416
Human_E	3113	0.0185	0.0242	0.0003	0.0097	0.1311
Human_A	3113	16.1950	0.6765	15.5610	15.9480	19.8252
Eregulat	3113	0.6787	0.5566	0.0000	0.4266	2.5853
Fsupervis	3113	0.0102	0.0109	0.0005	0.0064	0.1116
Gover	3113	0.2005	0.1021	0.0352	0.1746	0.8717
Ecopen	3113	0.1874	0.3045	0.0000	0.0770	4.6784
Urban	3113	0.5547	0.1472	0.0649	0.5290	0.9973
Fincial	3113	16.4050	1.1435	13.7230	16.1831	20.4202
IndStr	3113	0.4692	0.1096	0.1068	0.4754	0.8934
Indus	3113	0.4015	0.1215	0.03245	0.4056	0.8821
Popul	3113	4.3752	3.3979	0.0510	3.6530	26.4810
Green	3113	0.3942	0.0692	0.0036	0.4050	0.8925
Fdi	3113	0.0167	0.0179	0.0000	0.0114	0.2116
InfStra	3113	17.0810	7.2722	0.0005	15.6770	60.0705

4. Baseline Empirical Results and Economic Analysis

4.1. The Impact of Fintech on GTFP

Table 4 provides an empirical test of the benchmark "fintech–GTFP" relationship. The stepwise regression method was used to overcome the problem of multicollinearity. Columns (1), (3), and (5) are the OLS regression results without fixed effects. Columns (2), (4), and (6) control for the city- and year-fixed effects to mitigate omitted variable

interference. The regression results show, in columns (1) and (2), considering fixed effects and control variables, whether the regression coefficients of the fintech development index (L.Fintech_P) on GTFP were positive and significant at least at the 5% confidence level. This indicates that fintech development can have a significant driving effect on GTFP: the better the regional fintech development, the higher the city's GTFP. As the control variables were gradually introduced, the estimated coefficient of the effect of fintech remained significantly positive at the 1% confidence level, and the conclusion that fintech drives city GTFP remained unchanged. From an economic perspective, taking column (6) as an example (including control variables and two-way fixed effects), the coefficient of the fintech development index was 0.0227 and significant at the 1% level. For every 1% increase in the regional fintech development index, the city GTFP increased by an average of 0.023 units. These results support Hypothesis 1.

Table 4. Baseline regression results.

Variable	(1)	(2)	(3)	(4)	(5)	(6)
L.Fintech_P	0.0178 *** (4.4721)	0.0135 ** (2.2875)	0.0136 *** (5.9293)	0.0163 *** (3.4491)	0.0145 *** (4.2998)	0.0227 *** (3.9538)
Gover			−0.0004 (−0.0419)	−0.1852 * (−1.6261)	−0.0027 (−0.2322)	−0.1845 (−1.2969)
Ecopen			−0.0052 (−1.3136)	−0.0067 (−1.1708)	−0.0043 (−1.0344)	−0.0057 (−0.9971)
Urban			0.0197 (0.0956)	−0.0765 * (−1.7174)	−0.0196 (−0.8750)	−0.0699 (−1.5111)
Indus			−0.0161 (−0.6194)	−0.0096 (−0.1849)	−0.0143 (−0.5496)	−0.0093 (−0.1823)
Fincial			0.0043 * (1.6835)	0.0051 * (1.6161)	0.0043 *** (3.6791)	0.0042 ** (2.4300)
IndStr			−0.0334 * (−1.6477)	−0.0172 ** (−2.3612)	−0.0287 ** (−2.0615)	−0.0149 ** (−2.3183)
Popul					0.0000 (0.1174)	0.0028 (0.4134)
Green					0.0055 (0.2792)	0.0270 (0.8726)
Fdi					0.0134** (2.2407)	0.2512** (2.0268)
Infstr					0.0003 *** (2.8775)	0.0006** (2.2499)
Cons	0.9135 *** (17.2697)	0.9358 (17.4054)	0.8633 *** (18.7024)	1.0972 *** (7.6931)	0.8662 *** (10.4305)	1.0370 *** (7.2836)
City fixed	No	YES	No	YES	No	Yes
Year fixed	No	YES	No	YES	No	Yes
Adj. R^2	0.5021	0.61001	0.5945	0.7271	0.6952	0.6838
N	2830	2830	2830	2830	2830	2830

Note: Values in the lower brackets of the estimated coefficients are t-statistics adjusted for robust standard errors (clustering at the city level); ***, **, and * represent passing significance tests at the 1%, 5%, and 10% confidence levels; the dependent variables in the regressions in columns (1) to (6) are GTFP_GML.

The significance and signs of the other control variables' coefficients were generally consistent with the existing literature. For example, taking column (6), which considers both fixed effects, the coefficient of government size (Gover) was negative but insignificant, which may be explained by the fact that the non-market-oriented behavior of government intervention is detrimental to the optimal allocation of resources and has a certain

inhibitory effect on the increase of production efficiency. The estimated coefficients of urbanization level (Urban) and industrialization level (Indus) were also insignificantly negative, suggesting that rapid urbanization and industrial expansion without considering environmental carrying capacity may exacerbate the contradictions between humans and nature, thus having a negative effect on GTFP. The coefficient of industrial structure (Indstr) was significantly negative, suggesting that the current imbalance in industrial structure is an important factor hindering the progress of China's GTFP. The estimated coefficient of financial development (Fincial) was significantly positive, which indicates that effective financial support is indispensable in enhancing GTFP. The coefficient of trade openness (Ecopen) was significantly positive, implying that international trade is an important support for the continuous improvement of economic development quality and efficiency. Foreign direct investment (Fdi) significantly improved GTFP. Generally speaking, foreign direct investment entities are mostly multinational corporations with advanced business models and frontier technology, which can produce good demonstration effects, incentive effects, and competitive effects on host country enterprises [61], and have a significant improvement effect on GTFP [62].

We further examined the dynamic impact of fintech on GTFP. Following Chen Zhongfei and Jiang Kangqi (2021) [12], we replaced the core explanatory variable in model (1) with the lagged fintech development index, and the results are reported in panel A of Table 5. With the increase in the lagged order of fintech, the intensity of the positive effect of fintech on GTFP remained basically stable, and all of them were significant at the 1% confidence level. This indicates that the incentive effect of fintech on GTFP has a dynamic and stable trend over time, and can continue to provide strong support for GTFP.

To more accurately portray the impact of Fintech on GTFP, we decomposed the fintech development index into two symmetrical dimensions: the breadth and depth of fintech development. We explored the impact of the different dimensions of fintech development on GTFP. The breadth of fintech development was mainly characterized as the coverage of electronic accounts related to regional fintech services [63], denoted as Fintech_B. The depth of fintech development reflects the frequency and transaction volume of local actual use of fintech services, representing the service capability of fintech [13], which was denoted as Fintech_D. We replaced the core explanatory variable in model (1) with the breadth and depth of fintech development and the test results are shown in panels B and C of Table 5, respectively. It was found that the positive impact of fintech development breadth on GTFP gradually decayed with the increase of lag periods, and its significance also declined or even disappeared. Specifically, the regression coefficient of fintech at lag 2 was 0.013 and passes the 1% statistical significance test, but the positive effect at lag 3 was weaker and significant only at the 10% confidence level, and the regression coefficient at lag 4 further declined and failed the statistical significance test. In contrast, the effect of depth of fintech development on GTFP was very prominent and remained significant with increasing lag. Overall, the depth of fintech development maintained a significant promoting effect on GTFP in a relatively long time series, exhibiting obvious dynamic cumulative characteristics.

The above results indicate that although both the depth and breadth of fintech development are conducive to improving GTFP, the promotion effect of fintech development depth was even stronger. Therefore, in the construction of fintech, it is not only necessary to expand the breadth of fintech development and achieve inclusive finance but also to pay more attention to tapping the depth of fintech development to fully improve fintech service capability.

Table 5. Dynamic impact of fintech on GTFP.

Variable	GTFP_GML		
	(1)	(2)	(3)
Panel A Comprehensive Index of Fintech Development			
L2.Fintech_P	0.0206 *** (7.7008)		
L3.Fintech_P		0.0193 *** (7.7074)	
L4.Fintech_P			0.0211 *** (5.7433)
Controls	Yes	Yes	Yes
Year/City fixed	Yes	Yes	Yes
Adj. R^2	0.5208	0.4792	0.4654
N	2547	2264	1981
Panel B The Breadth of Fintech Development			
L2.Fintech_B	0.0130 *** (3.9277)		
L3.Fintech_B		0.0085 * (1.8621)	
L4.Fintech_B			0.0033 (0.9516)
Controls	Yes	Yes	Yes
Year/City fixed	Yes	Yes	Yes
Adj. R^2	0.3208	0.3784	0.3656
N	2547	2264	1981
Panel C The Depth of Fintech Development			
L2.Fintech_D	0.0255 ** (2.2856)		
L3.Fintech_D		0.0293 *** (3.9901)	
L4.Fintech_D			0.0313 *** (3.6856)
Controls	Yes	Yes	Yes
Year/City fixed	Yes	Yes	Yes
Adj. R^2	0.4208	0.4812	0.4665
N	2547	2264	1981

Note: Values in the lower brackets of the estimated coefficients are t-statistics adjusted for robust standard errors (clustering at the city level); ***, **, and * represent passing significance tests at the 1%, 5%, and 10% confidence levels.

4.2. Discussion and Treatment of Endogeneity

In the benchmark regression, we treated the fintech indicators with a one-period lag to minimize the impact of reverse causality on the empirical results, while controlling for factors that may simultaneously affect fintech and GTFP. However, the empirical results may have still been subject to unobservable factors or reverse causality, resulting in biased OLS estimates. To alleviate the endogeneity issues, this paper employed the instrumental variable method and GMM dynamic panel analysis.

4.2.1. Instrumental Variable Method

Drawing on the research of Xie Xuanli et al. (2018) [25] and Huang Qunhui et al. (2019) [64], the Internet penetration rate and the number of landline telephones per million people in each province of China were used as instrumental variables. The reasons are as follows: first, the Internet, as the infrastructure of fintech, is closely related to the changes in fintech. Second, the application of internet technology in China started from telephone penetration, and the development of fintech also originated from fixed telephone penetration, so regions with higher fixed telephone penetration may have a higher level of fintech development. Third, both the internet penetration rate in 1999 and the number of fixed telephones per 10,000 people in 1999 are historical data, which are obviously exogenous and have a small quantitative relationship with the current GTFP, and thus satisfy the principle of exclusivity and are suitable as instrumental variables. Since the selected instrumental variables are cross-sectional in form and cannot be directly used in panel data analysis, the final instrumental variables IV1 and IV2 were obtained by referring to Manacorda and Tesei (2020) [65] and using the time trend term reflecting the evolution of fintech multiplied by the internet penetration rate in 1999 and the number of fixed telephones per 10,000 people in 1999, respectively.

Columns (1) and (2) of Table 6 report the results of the two-stage least squares regression (TSLS) using the instrumental variables. In the first stage, the estimated coefficients of IV1 and IV2 were significant at the 1% confidence level, indicating that regions with more developed Internet and higher fixed telephone penetration rates historically also had better fintech development, consistent with the previous analysis. In the second stage, the regression coefficient of fintech was significantly positive, suggesting that after controlling for potential endogeneity, the development of fintech indeed provided a driving force for GTFP. Moreover, the coefficients of fintech were significantly higher than those in the benchmark regression in Table 4, which is consistent with previous instrumental variable estimation results. In a series of statistical tests for instrument validity, both the K-P rk LM and K-P Wald rk F tests significantly rejected the null hypothesis of under-identification and weak identification, indicating that the instrumental variables had a strong association with endogenous variables and were strong instrumental variables. The Hansen J test also failed to reject the null hypothesis of exogeneity of the instrumental variables at the 10% significance level, indicating that the instrumental variables were strictly exogenous.

4.2.2. GMM Dynamic Panel Analysis

The city GTFP may have a certain degree of persistence, i.e., serial correlation. To address this issue, the benchmark model (1) was extended as follows:

$$GTFP_{i,t} = \beta_0 + \beta_1 GTFP_{i,t-1} + \beta_2 Fintech_{m,t-1} + \sum_{k=3}^{T} \beta_k Controls_{i,t} + \mu_i + \delta_t + \varepsilon_{i,t} \quad (5)$$

The variable settings are consistent with the previous section. To overcome the possible endogeneity in the model and to reduce measurement errors, omitted variables, and other issues, the two-step system generalized method of moments (SYS-GMM) was employed to estimate the parameters of Equation (5). The results are shown in columns (3) and (4) of Table 6. The coefficients of fintech were significantly positive in both sets of regressions, indicating that the contribution of fintech to GTFP still exists when considering the serial correlation problem, and the previous conclusion is robust. In addition, the Sargan–Hansen test statistics failed to reject the null hypothesis, implying that the choice of instrumental variables was valid; the AR(1) test rejected the null hypothesis, while the AR(2) test accepted the null hypothesis, inferring that there was no serial correlation in the error terms of the original model, and thus the estimation results were not affected by the serial correlation of error terms.

Table 6. Endogeneity treatment results.

Panel A The Second-Stage Regression				
Variable	**TSLS**		**GMM Analysis**	
	(1)	**(2)**	**(3)**	**(4)**
L.GTFP			0.1073 ***	0.1140 ***
			(5.4649)	(6.0710)
L.Fintech_P	0.0632 ***	0.1403 ***	0.0449 ***	0.0309 ***
	(3.7659)	(3.6390)	(4.2432)	(2.8325)
Controls	No	Yes	No	Yes
City fixed	Yes	Yes	Yes	Yes
Year fixed	Yes	Yes	Yes	Yes
D-W-H Test	27.4403	22.1133		
(*p*-value)	(0.000)	(0.000)		
K-P rk LM	39.1123	24.1961		
(*p*-value)	(0.000)	(0.000)		
K-P wald rk F	70.9419	36.5875		
(*p*-value)	(0.000)	(0.000)		
Hansen/Sargan	2.9897	3.5917	9.8577	5.6650
(*p*-value)	(0.1184)	(0.1518)	(0.2326)	(0.1692)
AR(1)			4.2605	3.0339
(*p*-value)			(0.0000)	(0.0006)
AR(2)			0.5282	0.6413
(*p*-value)			(0.6041)	(0.5251)
Adj. R^2	0.4937	0.5546		
N	2830	2830	2547	2547
Panel B The First-Stage Regression				
IV	**(1)**	**(2)**	**(3)**	**(4)**
IV1	0.1461 ***	0.2079 ***		
	(3.3238)	(6.3869)		
IV2	0.1112 ***	0.1383 ***		
	(4.6138)	(3.7475)		
F	134.0743	95.1193		
(*p*-value)	(0.0000)	(0.0000)		

Note: D-W-H, K-P rk LM, K-P wald rk F, Hansen/Sargan represent endogeneity, under-identification, weak identification, and overidentification test statistics, respectively; AR(1) and AR(2) are the first-order and second-order serial correlation test statistics, respectively. The accompanying probabilities corresponding to each statistic are shown in the lower parentheses. *** represent passing significance tests at the 1% confidence levels.

4.3. Robustness Test

4.3.1. Replacing the Measurement Indicators of Fintech

First, there may be significant differences in economic growth, financial development, and other aspects among different cities within the same province, which may affect the conclusions of this paper. In view of this, we replaced the core explanatory variable in the benchmark model (1) with the city-level fintech development index (Fintech_C), and the test results are reported in column (1) of Table 7. As can be seen, the positive effect of city-level fintech on GTFP remained significant. The regression results are robust.

Second, following the approach of Song Min et al. (2020) [10] and considering both the financial function and financial technology characteristics, we used the logarithm of the number of provincial fintech companies as a proxy variable for fintech development (Fintech_N), replacing the core explanatory variable in the benchmark model (1) for re-estimation; the test results are reported in column (2) of Table 7. The conclusions still held.

Table 7. Robustness test results.

Variable	(1)	(2)	(3)	(4)	(5)
	Fintech_C	**Fintech_N**	**GTFP_DDF**	**Exclusion of Specific Samples**	**Winsorized**
L.Fintech_C	0.0185 *** (2.9372)				
L.Fintech_N		0.0117 ** (2.2582)			
L.Fintech_P			0.0420 *** (7.4359)	0.0347 *** (3.7553)	0.0179 *** (4.3942)
Controls	Yes	Yes	Yes	Yes	Yes
City fixed	Yes	Yes	Yes	Yes	Yes
Year fixed	Yes	Yes	Yes	Yes	Yes
Adj. R^2	0.5023	0.5041	0.5046	0.4459	0.4542
N	2830	2830	2830	2600	2830

Note: Values in the lower brackets of the estimated coefficients are *t*-statistics adjusted for robust standard errors (clustering at the city level); *** and ** represent passing significance tests at the 1%, and 5%, confidence levels.

4.3.2. Replacing the Explanatory Variable

To ensure that the benchmark regression results will not vary due to different measurement methods of GTFP, we further used the SBM directional distance function (SBM-DDF) including non-expected output to re-estimate the city-level green total factor productivity (GTFP_DDF), replacing the dependent variable in model (1) for re-testing to ensure the robustness of the benchmark results. The test results are shown in column (3) of Table 7. The results show that the benchmark results of this paper were still significantly valid.

4.3.3. Exclusion of Specific Samples

Municipalities and provincial capital cities have greater advantages in terms of fintech and GTFP, and the reverse causality problem may be more serious. In accordance with Chen Zhongfei and Jiang Kangqi (2021) [12], we excluded municipality and provincial capital city samples and re-estimated the benchmark model (1); the test results are reported in column (4) of Table 7. The results show that after excluding specific samples, the coefficient of fintech on GTFP was still significantly positive, which is consistent with the benchmark regression results.

4.3.4. Winsorize

To avoid the impact of large fluctuations in the dependent variable on the regression results, we winsorized the dependent variable at 1% and 99% levels and re-estimated the benchmark model. The results are shown in column (5) of Table 7. The test results are nearly identical to the benchmark regression results, indicating that the conclusions are reliable.

4.4. Further Analysis: Non-Linear Incentive Effect of Fintech on GTFP

As pointed out in the theoretical analysis, the impact of fintech on GTFP may exhibit heterogeneous characteristics depending on its development stage. To test this theoretical hypothesis, we constructed a dynamic panel threshold model with fintech development as the threshold variable. Considering that endogeneity may lead to biased estimation, we followed Seo and Shin (2016) [66] and adopted the first-order difference transformation generalized moment estimation method (FD-GMM), which can effectively overcome the joint endogeneity of the threshold variable and core explanatory variable. The dynamic panel threshold model was set as follows:

$$GTFP_{i,t} = \lambda_0 + \lambda_1 GTFP_{i,t-1} + \lambda_2 Fintech_{m,t-1}(Fintech_{m,t} \leq \omega) + \lambda_3 Fintech_{m,t-1}(Fintech_{m,t} > \omega) + \sum_{k=4}^{T} \lambda_k Controls_{i,t} + \mu_i + \delta_t + \varepsilon_{i,t} \quad (6)$$

Here, Fintech is both the core explanatory variable and threshold variable; ω is the threshold value to be estimated; $I(\cdot)$ is an indicator function, which takes the value of 1 when the condition in the parenthesis holds and 0 otherwise; λ is the coefficient term of the regression variable; μ_i represents the city fixed effect; θ_t represents the time fixed effect; and $\varepsilon_{i,t}$ is the random error, $\varepsilon_{i,t} \sim iid(0, \sigma^2)$. Equation (6) is a single-threshold case and can be expanded to a multi-threshold case based on sample data statistical tests and other steps.

The results of the bootstrap test and threshold value estimation are shown in panel A and panel B of Table 8. It can be observed that the fintech development index passed the double-threshold effect test, indicating that the impact of fintech on GTFP will exhibit heterogeneity as its development stage changes. Using the FD-GMM approach to estimate the dynamic panel threshold model can overcome the adverse effects of endogeneity to some extent, and the estimation results are shown in panel C of Table 8. With the development of fintech, the intensity of the positive impact of fintech on GTFP rose continuously. Specifically, when Fintech $\leq$ 5.6909, the estimated coefficient of Fintech was 0.0018 at the 5% level; when 5.6909< Fintech $\leq$ 5.8000, the coefficient of Fintech was 0.0090 and significant at the 5% level; when Fintech $>$ 5.8000, the estimated coefficient became 0.0142 and passed the 1% statistical significance test. Thus, theoretical Hypothesis 3 is confirmed.

Table 8. Dynamic panel threshold effect tests and parameter estimation.

Panel A Threshold Effect Test Results							
Variable	Threshold Effect	SupW	BS	*p*-Value	Critical Value		
					1%	5%	10%
Fintech_P	Single	80.0305 ***	1000	0.000	16.8911	12.6236	11.6284
	Double	54.8497 ***	1000	0.007	24.6122	13.6262	11.5559
	Triple	24.4906	1000	0.713	109.8550	92.7709	86.3673
Panel B Threshold Estimation Results							
Threshold Type	Threshold Value 1				Threshold Value 2		
	Threshold Estimate		95% Confidence Interval		Threshold Estimate	95% Confidence Interval	
Fintech_P	5.6909		[5.6902, 5.6911]		5.8000	[5.7911, 5.8049]	
Panel C Estimation Results of Dynamic Panel Threshold Model							
Variable	**Coefficient**					***p*-Value**	
L.GTFP	0.1902 *** (4.0637)					0.000	
$I(Fintech_P > \omega_2)$	0.0142 *** (4.9505)					0.000	
$I(\omega_1 < Fintech_P \leq \omega_2)$	0.0090 ** (2.2511)					0.025	
$I(Fintech_P \leq \omega_1)$	0.0018 ** (2.0345)					0.044	
Controls	Yes						
Hansen J (*p*-value)	47.0746 (0.151)						
Wald	9638 ***						
N	2547						

Note: **, and *** indicate significance at the 5%, and 1% confidence levels, respectively. Values in the lower brackets of the estimated coefficients are the corresponding z-statistics. L.GTFP is the first-order lag of the explained variable. The Hansen J is a test statistic for overfitting of instrumental variables, with the null hypothesis being that the instrumental variables are valid, and its corresponding *p*-value is in the parentheses below.

5. Identification Test of the Mechanism of Fintech Affecting GTFP

The theoretical analysis section mentioned that fintech development can drive GTFP through the channels of resource allocation, human capital, and technological innovation. Based on the available data, we then conducted an empirical test of the above mechanisms. Following Jiang Ting's (2022) operational suggestions for analyzing mediation effects [67], the mechanism test mainly focused on identifying the effectiveness of fintech on mediating variables. The following econometric model was constructed:

$$\text{Mediator}_{i,t} = \eta_0 + \eta_1 \text{Fintech}_{m,t-1} + \sum_{k=2}^{T} \eta_k \text{controls}_{i,t-1} + \delta_t + \mu_i + \varepsilon_{i,t} \quad (7)$$

On the selection of mediating variables (Mediator), resource allocation efficiency proxies included the capital mismatch index (Capital_M) and labor mismatch index (Labor_M); human capital proxies included the human capital accumulation level (Human_E) and human capital structural upgrading (Human_A); and technology innovation proxies included technology innovation quantity (Innova_N) and technology innovation quality (Innova_Q). The other variables were set as described above. The detailed test results are shown in Table 8.

From columns (1)–(2) of Table 9, the regression coefficients of fintech on the capital mismatch index and the labor mismatch index were −0.0896 and −0.1145, respectively, and both were significant at least at the 5% confidence level, indicating that fintech development can effectively reduce the degree of capital and labor mismatch and produces a significant improvement in resource allocation. Hsieh and Klenow (2009) [27] believe that focusing on resource allocation efficiency is essential for enhancing TFP in a transitioning economy like China's. Relying on information technology, fintech can facilitate the effective transmission of market information, guide the allocation of factor resources across time and space, improve the efficiency of resource circulation, and greatly facilitate the exchange of factors between regions and industries, thus promoting the overall progress of GTFP.

Table 9. Mechanism test results.

Variable	(1)	(2)	(3)	(4)	(5)	(6)
	Capital_M	**Labor_M**	**Human_E**	**Human_A**	**Innova_N**	**Innoa_Q**
L.Fintech_P	−0.0896 ** (−2.3794)	−0.1145 *** (−2.8944)	0.0035 ** (2.5043)	0.4072 *** (3.9092)	0.1848 * (1.7179)	0.4207 *** (3.8758)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
City fixed	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed	Yes	Yes	Yes	Yes	Yes	Yes
N	2830	2830	2830	2830	2830	2830
Adj. R^2	0.8397	0.7539	0.9599	0.9598	0.9740	0.8345

Note: Values in the lower brackets of the estimated coefficients are *t*-statistics adjusted for robust standard errors (clustering at the city level); ***, **, and * represent passing significance tests at the 1%, 5%, and 10% confidence levels.

According to the test results of columns (3) and (4) in Table 9, fintech had a significant growth-promoting effect on human capital accumulation and human capital structure upgrading, which lays a solid talent foundation for promoting environmental governance, cultivating green consumption concepts, and promoting green production, thus improving GTFP. Since human capital is an important determinant of technological innovation, fintech should have a certain "innovation-promoting" effect following the above logic. According to columns (5) and (6) of Table 9, fintech had a significant promoting effect on technological innovation (coefficients were 0.1848 and 0.4207, respectively), providing empirical support for the above speculation. Specifically, the pulling effect of fintech development on the quality of technological innovation was particularly strong. The possible reasons include:

First, fintech can enhance the ability of entities to collect, integrate, and analyze information, help them to grasp the situation of technological innovation and market potential in a timely manner, and improve the effectiveness of technological innovation decisions. Second, the development of fintech is conducive to improving market competition mechanisms [28]; under external market supervision pressure, enterprises will focus more on enhancing their core competitiveness. Both domestic and foreign researchers believe that technological innovation is crucial for improving green total factor productivity [5,20,29,60]. Therefore, financial technology can help improve GTFP by driving technological innovation.

In summary, improving resource allocation efficiency, optimizing human capital, and promoting technological innovation are important channels for fintech to drive GTFP. The empirical results support Hypothesis 2.

6. Analysis of the Moderating Effect of Fintech on GTFP

6.1. Linear Moderating Effect

The theoretical analysis mentioned that the driving effect of fintech on GTFP may be subject to the moderating influence of financial regulation and environmental regulation. Therefore, this section further explores the "fintech–GTFP" research paradigm by embedding financial regulation and environmental regulation elements. First, the linear moderating effect was examined, and the econometric model was set as follows:

$$GTFP_{i,t} = \xi_0' + \xi_1' Fintech_{m,t-1} + \xi_2' Adjuva_{m,t-1} + \xi_3' \left(Fintech_{m,t-1} \times Adjuva_{m,t-1} \right) + \sum_{k=4}^{T} \xi_k' Controls_{i,t} + \mu_i + \delta_t + \varepsilon_{i,t} \quad (8)$$

In Equation (8), Adjuva represents the province-level moderating variables, including financial regulation (Fsupervis) and environmental regulation (Eregulat). The meanings of the other variables are consistent with the previous sections. We focused on the coefficient signs of Fintech × Adjuva to reveal the moderating effect of financial regulation and environmental regulation on the baseline relationship between fintech and GTFP.

The regression results of Equation (8) are reported in Table 10. It was found that the cross-product term coefficient of financial regulation and fintech was significantly positive, indicating that better financial regulation is an important condition for fintech to promote GTFP. Financial regulation is conducive to reducing the financial risks arising from fintech innovation, ensuring the standardized and orderly development of fintech, and better realizing the important goal of "financial support for sustainable development of real economy". Meanwhile, the cross-product of environmental regulation and fintech had a weak contribution to GTFP (coefficient was 0.0022) and did not pass the significance test. On the one hand, this result suggests that environmental regulation does enhance the contribution of fintech to GTFP to a certain extent; on the other hand, it also reminds us that the "innovation compensation" and "cost compliance" attributes of environmental regulation itself may determine that its moderating effect is not a simple linear one [68]. If a linear model is used for fitting, it may obscure part of the truth. However, the above results preliminarily verify Hypothesis 4.

Table 10. Results of the test for linear moderation effect.

Variable	(1)	(2)
	Fsupervis	**Eregulat**
L.Fintech_P	0.0222 ** (2.2348)	0.0221 * (1.8207)
L.Fsupervis	0.0483 (0.0847)	
L.Eregulat		−0.0821 *** (−4.7927)
L.Fintech_P × L.Fsupervis	0.0111 ** (2.3615)	
L.Fintech_P × L.Eregulat		0.0022 (0.2845)
Controls	Yes	Yes
City fixed	Yes	Yes
Year fixed	Yes	Yes
N	2830	2830
Adj. R^2	0.3034	0.3141

Note: Values in the lower brackets of the estimated coefficients are *t*-statistics adjusted for robust standard errors (clustering at the city level); ***, **, and * represent passing significance tests at the 1%, 5%, and 10% confidence levels.

6.2. Nonlinear Moderating Effect

The empirical test in the previous section confirmed that financial regulation and environmental regulation have a certain moderating strengthening effect on the benchmark relationship between fintech and GTFP. However, do different intervals of financial regulation and environmental regulation make a large difference in the moderating effect and lead to a non-linear statistical characteristic of the impact of fintech on GTFP? To answer this question, we used a dynamic panel threshold model to explore the nonlinear relationship between fintech and GTFP from the perspectives of financial regulation and environmental regulation. Considering that the joint endogeneity of the core explanatory variable (fintech) and the threshold variable (financial regulation, environmental regulation) may cause estimation bias, consistent with the previous sections, the FD-GMM method that can overcome the joint endogeneity was adopted. The dynamic panel threshold model was set as follows:

$$GTFP_{i,t} = \kappa_0 + \kappa_1 GTFP_{i,t-1} + \kappa_2 Fintech_{m,t-1}\left(Adjuva_{m,t} \le v\right) + \kappa_3 Fintech_{m,t-1}\left(Adjuva_{m,t} > v\right) + \sum_{k=4}^{T} \kappa_k Controls_{i,t} + \mu_i + \delta_t + \varepsilon_{i,t} \quad (9)$$

where Adjuva is the threshold variable, including financial regulation (Fsupervis) and environmental regulation (Eregulat); fintech (Fintech) is the core explanatory variable; $I(\cdot)$ is an indicator function, which takes a value of 1 when the condition in the parentheses holds and 0 otherwise; and κ is the coefficient term of the regression variable. The other variables are consistent with the previous sections.

Using financial regulation or environmental regulation as the threshold variable, we conducted bootstrap sampling tests to determine whether threshold effects exist between fintech and GTFP. The test results are shown in panel A of Table 11. Both financial regulation and environmental regulation passed the double-threshold effect test; therefore, the double-threshold model is applicable, which also indicates that a non-linear structural transformation effect of the impact of fintech on GTFP exists based on these two threshold variables. Panel B of Table 11 further shows the threshold estimates and 5% confidence intervals for the two threshold variables.

Table 11. Threshold effect tests and threshold estimates.

Panel A Threshold effect test results							
Variable	Threshold Effect	SupW	BS	*p*-Value	Critical Value		
					1%	5%	10%
Fsupervis	Single	17.0446	1000	0.000	13.6705	10.1751	8.3059
	Double	24.0863	1000	0.000	12.7142	11.0066	9.0383
	Triple	14.4504	1000	0.420	39.1392	31.7894	28.0710
Eregulat	Single	29.4466	1000	0.000	19.7083	16.6235	14.4088
	Double	24.6195	1000	0.000	18.5559	15.0787	11.0093
	Triple	15.1822	1000	0.700	85.0688	76.1750	60.5853

Panel B Threshold estimation results				
Threshold Type	Threshold Value 1		Threshold Value 2	
	Threshold Estimate	95% Confidence Interval	Threshold Estimate	95% Confidence Interval
Fsupervis	0.0136	[0.0129, 0.0159]	0.0461	[0.0460, 0.0463]
Eregulat	0.4183	[0.4050, 0.4187]	1.4354	[1.4292, 1.4363]

Next, we conducted an overall estimation of the dynamic panel threshold model; the detailed results are reported in Table 12. Column (1) shows the regression results with financial regulation as the threshold variable. Financial regulation has a significant double-threshold characteristic, and the impact strength of fintech on GTFP varied significantly within different financial regulation zones. When financial regulation was below the threshold value of 0.0136, the coefficient value of fintech was 0.0144, which is weak; when financial regulation was between (0.0136, 0.0461), the regression coefficient of fintech rose to 0.1197 and was significant at the 1% level; when financial regulation crossed the maximum threshold value of 0.0461, the impact effect of fintech on GTFP remained positive but was significantly weakened and no longer significant. It can be seen that there is an optimal zone for the moderating effect of financial regulation. Weakened financial regulation may not be able to suppress the financial fluctuations or risk shocks caused by financial technology innovation, making it difficult for fintech to achieve its intended economic effects. Excessive financial regulation may weaken the vitality of the financial market, constrain financial technology innovation, distort the allocation of financial resources, and fail to leverage the function of fintech to improve economic quality and efficiency. Therefore, maintaining moderate and effective financial regulation is preferable.

Similarly, the regression results with environmental regulation as the threshold variable shown in Column (2) are also mixed, and the impact of fintech on GTFP exhibited significant heterogeneity within different environmental regulation zones. When the environmental regulation was less than 0.4183, the coefficient of fintech was 0.0620 and significant at the 10% confidence level; when the environmental regulation was in the (0.4183, 1.4354) interval, its coefficient was 0.1044 at the 1% significance level; and when the environmental regulation exceeded 1.4354, its coefficient was −0.0168 at the 10% significance level. It was easy to find that there was also an optimal range for the moderating effect of environmental regulation. The government's adoption of moderate environmental regulation policies will have a "forcing effect" and a "guiding effect" on enterprises, allowing enterprises benefiting from fintech to have enough time and energy for production equipment upgrades and increasing green technology R&D, ultimately reducing pollution emissions during the production process, thereby enhancing the level of GTFP. However, overly stringent environmental control will make economic entities exhausted in fulfilling their environmental protection responsibilities and neglect production and technological innovation, the internal and external environment for the stable operation of enterprises

will face more uncertainties, and the support effect of fintech on enterprises will naturally be greatly discounted, thereby hindering the progress of GTFP.

Table 12. Estimation Results of Dynamic Panel Threshold Model.

Variable	(1)		(2)	
	Fsupervis		Eregulat	
	Coefficient	*p*-Value	Coefficient	*p*-Value
L.GTFP	0.1480 *** (3.2323)	0.000	0.1127 *** (4.0226)	0.000
$I(Adjuva > v_2)$	0.0146 (1.5941)	0.110	−0.0168 * (1.8914)	0.058
$I(v_1 < Adjuva \leq v_2)$	0.1197 *** (5.4812)	0.000	0.1044 *** (4.3907)	0.000
$I(Adjuva \leq v_1)$	0.0144 ** (2.0630)	0.040	0.0620 * (1.8144)	0.070
Controls	Yes		Yes	
Hansen J (*p*-value)	22.4656 (0.309)		17.7935 (0.449)	
Wald	11204 ***		9752 ***	
N	2547		2547	

Note: *, **, and *** indicate significance at the 10%, 5%, and 1% confidence levels, respectively. Values in the lower brackets of the estimated coefficients are the corresponding z-statistics. L.GTFP is the first-order lag of the explained variable. The Hansen J is a test statistic for overfitting of instrumental variables, with the null hypothesis being that the instrumental variables are valid, and its corresponding *p*-value is in the parentheses below.

In summary, the aforementioned results fully demonstrate that fintech has nonlinear threshold effects on GTFP based on financial regulation and environmental regulation. Financial regulation or environmental regulation that is too strong or too weak is not conducive to fully exploiting the promoting effect of fintech on GTFP, and only moderate financial regulation and environmental regulation policies can maximize the moderating strengthening effect of both. Thus, Hypothesis 4 is further verified. In addition, the Hansen J statistic measuring instrumental variable overfitting did not reject the null hypothesis of effective instrument variables, proving the validity of the instrument variables and the rationality of the model setting. The Wald statistic also showed that all dynamic panel threshold model estimation results were highly significant.

7. Research Conclusions and Policy Implications

7.1. Research Conclusions

Enhancing the ability of financial services to serve the real economy is crucial for achieving green transformation and high-quality development of the real economy. In the context of the continuous integration of emerging AI technology and the financial industry, exploring how the development of fintech affects GTFP undoubtedly has important theoretical value and practical significance. This paper first analyzed the relationship between fintech and GTFP from a theoretical perspective and then used the super-efficiency SBM model with embedded non-expected output and the globalization GML index method to measure the green total factor productivity of China's 283 prefecture-level and above cities from 2011 to 2021. On this basis, we empirically tested the impact of fintech on GTFP and its underlying mechanisms in a multi-dimensional manner. The main conclusions are as follows: First, the development of fintech significantly improved the level of city GTFP, and the impact effect showed dynamic stability characteristics. After endogeneity treatment and other robustness tests, the findings of the study still held.

Second, by decomposing fintech into two dimensions, depth and breadth of fintech development, it was found that both the depth and breadth of fintech development con-

tributed to the improvement of city GTFP. However, the depth of fintech development had a stronger and dynamically cumulative effect on city GTFP, while the breadth of fintech development had a relatively weaker and dynamically decaying effect. Specifically, in the more mature stages of fintech development, fintech will have a stronger and more significant incentive effect on city GTFP.

Third, the mechanism analysis showed that financial technology promoted the progress of city GTFP through channels such as improving resource allocation efficiency, optimizing human capital, and stimulating technological innovation.

Fourth, the moderating effect test showed that financial regulation and environmental regulation had a positive linear moderating effect on the benchmark relationship between fintech and GTFP. Further research revealed that the moderating effects of financial regulation and environmental regulation have significant nonlinear threshold characteristics, and when the intensity of financial regulation and environmental regulation was too high or too low, they could not fully exert their moderating effects; only when both are in the optimal range can the driving effect of fintech on GTFP be maximized.

7.2. Policy Implications

The findings of this paper provide important policy implications for exploring the drivers of GTFP growth and thus achieving high-quality economic development.

First, given the reality that fintech can play a constructive role in improving GTFP, the Chinese government should provide policies that support the deep integration of AI technology and finance, and promote the further development of the fintech industry in a comprehensive and multifaceted manner while keeping the bottom line of risk. On the one hand, policies should encourage and improve the development of supporting industries, deepen the application of AI technology in finance, establish a balanced system of safety and efficiency for the application of technological achievements, strengthen the protection of fintech patents, continuously expand the open innovation and win–win industrial ecology, enhance the market competition efficiency of the fintech industry, and improve the survival of the fittest mechanisms. On the other hand, in the development of fintech, efforts should be made to actively expand the coverage of fintech and focus on using AI technological advantages to improve the accessibility of financial services. In addition, we should vigorously promote the in-depth construction of fintech, and comprehensively improve the quality and efficiency of financial services through technology-driven and data-empowered fintech. Meanwhile, the government should comprehensively strengthen the construction of software and hardware infrastructure to consolidate the foundation of fintech development. Specifically, the Chinese government should accelerate the intelligent upgrading and transformation of the fintech infrastructure and the cultivation of high-level scientific and technological talents, focus on promoting the research and application of key AI technologies such as big data, cloud computing, and blockchain, and assist in deepening fintech capabilities by improving the construction level of AI infrastructure, laying a solid foundation for fintech to effectively contribute to the transformation and upgrading of the real economy.

Second, policies should vigorously promote the coordinated development of fintech and green industries, and build a healthy application ecosystem of fintech to support improvement of GTFP. On the one hand, the government should encourage financial institutions to fully utilize advanced AI technologies such as big data, cloud computing, and block chain to identify green industries (e.g., clean technology, clean production models, and clean enterprises), reduce the pre-approval procedures and post-supervision costs of green credit, increase funding support for the R&D and promotion of new green and energy-saving technologies and products, and ultimately improve the level of GTFP by focusing on green technology innovation and industrial structure upgrading. On the other hand, utilizing the information identification, analysis, and supervision capabilities of fintech, the government can help micro-enterprises grasp the situation in terms of technological innovation and market potential, enhance the effectiveness of enterprise

innovation decision-making and fund utilization, optimize resource allocation, and create favorable conditions for the progress of GTFP.

Third, this paper found that fintech development had a more significant contribution to GTFP in regions with moderate financial and environmental regulations. Therefore, the following suggestions are proposed: (1) Accelerate the reform of the financial supervision system, appropriately adjust the regulatory framework and improve relevant policies, and enhance the quality and efficiency of financial supervision services for fintech to meet the needs of the rapid development of fintech. At the same time, regulatory authorities should implement incentive-compatible regulatory principles, adopt sustainable, coordinated, and moderate regulatory policies, focus on process supervision, strengthen ex-ante and ex-post supervision, and improve the pertinence, timeliness, and penetration of supervision, building a firewall between finance and technology. (2) The government should grasp the intensity of environmental regulation, formulate appropriate environmental regulation policies, and avoid "campaign-style" law enforcement. Neither should excessive environmental control be imposed, nor should environmental governance be overly relaxed. Attention should be paid to the issue of "degree" in the process of formulating environmental standards, finding appropriate environmental governance methods, effectively urging enterprises to undergo green technology transformation and innovation, and enhancing the promoting effect of fintech on GTFP.

Author Contributions: Conceptualization, W.H. and X.L.; methodology, X.L.; software, W.H.; validation, X.L. and W.H.; formal analysis, W.H.; investigation, X.L.; resources, X.L.; data curation, W.H.; writing—original draft preparation, W.H.; writing—review and editing, X.L.; visualization, X.L.; supervision, W.H.; project administration, W.H.; funding acquisition, W.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received was funded by The National Social Science Fund of China, grant number 22BJY168, and National Social Science Fund Youth Project, grant number 20CJY065.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Song, D.; Zhang, Q. Evolution and Driving Forces of the Integration of Environmental Protection and Economic High-Quality Development. *J. Quant. Tech. Econ.* **2022**, *39*, 42–59. [CrossRef]
2. Niu, H.; Yan, C. Environmental Taxation, Resource Allocation, and Economic High-Quality Development. *World Econ.* **2021**, *44*, 28–50. [CrossRef]
3. Xiao, W.; Xue, T. Rising Labor Costs, Financing Constraints, and Total Factor Productivity Changes of Enterprises. *World Econ.* **2019**, *42*, 76–94. [CrossRef]
4. Jiang, L.; Chen, X.; Jiang, Y.; Zhang, B. Exploring the Direct and Spillover Effects of Aging on Green Total Factor Productivity in China: A Spatial Econometric Approach. *Sustainability* **2023**, *15*, 6709. [CrossRef]
5. He, L.; Qi, X. Environmental Regulation and Green Total Factor Productivity-Evidence from Chinese Industrial Enterprises. *Econ. Perspect.* **2022**, *736*, 97–114.
6. Zhou, Y.; Xu, Y.; Liu, C.; Fang, Z.; Fu, X.; He, M. The Threshold Effect of China's Financial Development on Green Total Factor Productivity. *Sustainability* **2019**, *11*, 3776. [CrossRef]
7. Liu, Z.; Ling, Y. Structural Transformation, Total Factor Productivity, and High-Quality Development. *Manag. World* **2020**, *36*, 15–29. [CrossRef]
8. Cai, W. The impact of banking market structure on firm productivity-empirical evidence from industrial firms. *Financ. Res.* **2019**, *2019*, 39–55.
9. Tang, S.; Lai, X.; Huang, R. How does financial technology innovation affect total factor productivity: Promotion or inhibition?—Theoretical analysis framework and regional practice. *China Soft. Sci.* **2019**, *343*, 134–144.
10. Song, M.; Zhou, P.; Si, H. Fintech and Total Factor Productivity of Firms-An "Empowerment" and Credit Rationing Perspective. *China Ind. Econ.* **2021**, *2021*, 138–155. [CrossRef]
11. Sheng, T.; Fan, C. Fintech, optimal banking market structure and credit supply for micro and small enterprises. *Financ. Res.* **2020**, *2020*, 114–132.

12. Chen, Z.; Jiang, K. Digital Financial Development and Enterprise Total Factor Productivity. *Econ. Perspect.* **2021**, *728*, 82–99.
13. Tang, S.; Wu, X.; Zhu, J. Digital finance and corporate technology innovation-structural characteristics, mechanism identification and differences in effects under financial regulation. *Manag. World* **2020**, *36*, 52–66. [CrossRef]
14. Zhu, T. A comprehensive analytical framework for the evolution of Fintech development in China. *Financ. Regul. Res.* **2018**, *2018*, 55–67. [CrossRef]
15. Hou, C.; Li, B. Does fintech increase total factor productivity-empirical evidence from Peking University's Digital Inclusive Finance Index. *Financ. Econ. Sci.* **2020**, *12*, 1–12.
16. Wang, X. Internet finance helps to solve the problem of difficult financing for "long-tail" small and micro enterprises. *Financ. Res.* **2015**, *2015*, 128–139.
17. Bunea, S.; Kogan, B.; Stolin, D. Banks versus fintech: At last, it's official. *J. Financ. Transform.* **2016**, *44*, 1–25.
18. Wang, J.; Zhang, Y.; Ma, X. Digital economy, resource mismatch and total factor productivity. *Financ. Trade Res.* **2022**, *33*, 10–26. [CrossRef]
19. Acemoglu, D.; Restrepo, P. Robots and Jobs: Evidence from US Labor Markets. *J. Political Econ.* **2020**, *128*, 2188–2244. [CrossRef]
20. Zhang, F.; Shi, Z.; Wu, G. The Impact of Digital Economy and Environmental Regulation on Green Total Factor Productivity. *Nanjing Soc. Sci.* **2022**, *28*, 12–20. [CrossRef]
21. ShangGuan, X.; Ge, B. Digital Finance, Environmental Regulation, and High-Quality Economic Development. *Mod. Financ. Econ. J. Tianjin Univ. Financ. Econ.* **2021**, *41*, 84–98. [CrossRef]
22. Cheng, W.; Qian, X. Digital Economy and Green Total Factor Productivity Growth of China's Industry. *Econ. Issues Explor.* **2021**, *2021*, 124–140.
23. Ma, G.; Lv, D.; Luo, Y.; Jiang, T. Environmental Regulation, Urban-Rural Income Gap and Agricultural Green Total Factor Productivity. *Sustainability* **2022**, *14*, 8995. [CrossRef]
24. Hua, S.; Li, J. Can Environmental Regulatory Tools Promote Green Technological Innovation for Enterprises under the Digital Economy Conditions? *Sci. Technol. Prog. Policy* **2023**, *40*, 141–150. [CrossRef]
25. Wang, X.; Song, M.; Yang, Y. Fintech, Financial Regulation, and High-Quality Development of Enterprises. *Res. Financ. Econ. Issues* **2023**, *473*, 87–99. [CrossRef]
26. Tone, K. A slacks-based measure of efficiency in data envelopment analysis. *Eur. J. Oper. Res.* **2001**, *130*, 498–509. [CrossRef]
27. Wang, Z.; Han, C.; Zhu, W. The Impact of Digital Financial Development on Export Technological Complexity. *World Econ. Stud.* **2022**, *342*, 26–42. [CrossRef]
28. Fuster, A.; Plosser, M.; Schnabl, P.; Vickery, J. The Role of Technology in Mortgage Lending. *Rev. Financ. Stud.* **2019**, *32*, 1854–1899. [CrossRef]
29. Acemoglu, D.; Aghion, P.; Bursztyn, L.; Hemous, D. The Environment and Directed Technical Change. *Am. Econ. Rev.* **2012**, *102*, 131–166. [CrossRef]
30. Demertzis, M.; Merler, S.; Wolff, G. Capital markets union and the fintech opportunity. *J. Financ. Regul.* **2018**, *4*, 157–165. [CrossRef]
31. Xie, X.; Shen, Y.; Zhang, H.; Guo, F. Can Digital Finance Promote Entrepreneurship?—Evidence from China. *Economics* **2018**, *17*, 1557–1580. [CrossRef]
32. Lee, H.; Yang, S.; Kim, K. *The Role of Fintech in Mitigating Information Friction in Supply Chain Finance*; ADB Economics Working Paper Series; ADB Publications: Manila, Philippines, 2019.
33. Hopenhayn, H. Firms, Misallocation, and Aggregate Productivity: A Review. *Soc. Sci. Electron. Publ.* **2014**, *6*, 735–770. [CrossRef]
34. Heiskanen, A. The technology of trust: How the Internet of things and blockchain could usher in a new era of construction productivity. *Constr. Res. Innov.* **2017**, *8*, 66–73. [CrossRef]
35. Shen, Y.; Guo, P. Internet finance, technology spillovers and total factor productivity of commercial banks. *Financ. Res.* **2015**, *2015*, 160–175.
36. Ba, S.; Bai, H.; Hu, W. Financial technology innovation, total factor productivity of firms and economic growth-based on a new structural economics perspective. *Res. Financ. Econ.* **2020**, *2020*, 46–53. [CrossRef]
37. Huang, Y.; Huang, Z. Digital financial development in China: Present and future. *Economics* **2018**, *17*, 1489–1502. [CrossRef]
38. Hsieh, C.; Klenow, P. Misallocation and Manufacturing TFP in China and India. *Q. J. Econ.* **2009**, *124*, 1403–1448. [CrossRef]
39. Pei, C.; Ni, J.; Li, Y. Analysis of the political economy of digital economy. *Financ. Trade Econ.* **2018**, *39*, 5–22. [CrossRef]
40. Xu, Z.; Zheng, F.; Chen, J. The "digital divide" or the "information dividend"? Effective supply of information and farmers' selling prices-an empirical study from a micro perspective. *Economics* **2013**, *12*, 1513–1536. [CrossRef]
41. Yang, F.; Huang, Y. Corporate governance structure, information asymmetry and SME financing. *Financ. Res.* **2006**, *2006*, 159–166.
42. Niu, Z.; Xu, C.; Wu, Y. Business environment optimization, human capital effect and firm labor productivity. *Manag. World* **2023**, *39*, 83–100. [CrossRef]
43. Mankiw, N.; Romer, D.; Weil, D. A contribution to the empirics of economic growth. *Q. J. Econ.* **1992**, *107*, 407–437. [CrossRef]
44. Daron, A.; Pascual, R. The race between man andmachine: Implications of technology for growth, factor shares, and employment. *Am. Econ. Rev.* **2018**, *108*, 1488–1542.
45. Galor, O.; David, W. Population, Technology, and Growth: From Malthusian Stagnation to the Demographic Transition and beyond. *Am. Econ. Rev.* **2000**, *90*, 806–828. [CrossRef]

46. Xiao, T.; Sun, R.; Yuan, C.; Sun, J. Enterprise Digital Transformation, Adjustment of Human Capital Structure, and Labor Income Share. *Manag. World* **2022**, *38*, 220–237. [CrossRef]
47. Zhao, T.; Zhang, Z.; Liang, S. Digital economy, entrepreneurial activity and high quality development-empirical evidence from Chinese cities. *Manag. World* **2020**, *36*, 65–76. [CrossRef]
48. Shang, X. A study on the mechanism of environmental regulation affecting the sustainability of economic growth-an empirical analysis based on the mediating effect. *J. Yunnan Univ. Financ. Econ.* **2022**, *38*, 1–17. [CrossRef]
49. Zhang, Y.; Kou, P. Environmental regulation, Internet penetration and corporate pollution emissions. *Ind. Econ. Rev.* **2018**, *9*, 128–139. [CrossRef]
50. Ambec, S.; Cohen, M.; Elgie, S.; Lanoie, P. The Porter Hypothesis at 20: Can Environmental Regulation Enhance Innovation and Competitiveness? *Rev. Environ. Econ. Policy* **2013**, *7*, 2–22. [CrossRef]
51. Porter, M.; Linde, C. Toward a New Conception of the Environment-Competitiveness Relationship. *J. Econ. Perspect.* **1995**, *9*, 97–118. [CrossRef]
52. Färe, R.; Grosskopf, S.; Pasurka, C.A., Jr. Environmental production functions and environmental directional distance functions. *Energy* **2007**, *32*, 1055–1066. [CrossRef]
53. Yu, Y. Re-estimation of inter-provincial total factor productivity in China from a heterogeneity perspective: 1978–2012. *Economics* **2017**, *16*, 1051–1072. [CrossRef]
54. Oh, D. A global Malmquist-Luenberger productivity index. *J. Product. Anal.* **2010**, *34*, 183–197. [CrossRef]
55. Fare, R.; Shawna, G. A comment on weak disposability in nonparametric production analysis. *Am. J. Agric. Econ.* **2009**, *91*, 535–538. [CrossRef]
56. Shan, H. Re-estimation of capital stock K in China: 1952 to 2006. *Res. Quant. Technol. Econ.* **2008**, *25*, 17–31.
57. Qiu, H.; Huang, Y.; Ji, Y. The Impact of Fintech on Traditional Banking Behavior-Based on the Perspective of Internet Wealth Management. *Financ. Res.* **2018**, *2018*, 17–29.
58. Meng, N.; Su, Q.; Lei, H. How financial technology affects competition in the banking industry. *Financ. Trade Econ.* **2020**, *41*, 66–79. [CrossRef]
59. Chen, Y.; Hu, W. Price distortions, factor mismatches and efficiency losses: Theory and applications. *Economics* **2011**, *10*, 1401–1422. [CrossRef]
60. Liu, Z.; Li, H.; Hu, Y.; Li, C. Advanced human capital structure and economic growth. *Econ. Res.* **2018**, *53*, 50–63.
61. Mukoyama, T. Innovation, imitation, and growth with cumulative technology. *J. Monet. Econ.* **2004**, *50*, 361–380. [CrossRef]
62. Fariborz, M.; Tian, X.; Zhang, B.; Zhang, W. Stock Market Liberalization and Innovation. *J. Financ. Econ.* **2020**, *8*, 985–1014. [CrossRef]
63. Li, C.; Yan, X.; Song, M.; Yang, W. Fintech and corporate innovation—Evidence from New Third Board listed companies. *China Ind. Econ.* **2020**, *2020*, 81–98. [CrossRef]
64. Huang, Q.; Yu, Y.; Zhang, S. Internet development and manufacturing productivity improvement: Intrinsic mechanisms and China's experience. *China Ind. Econ.* **2019**, *377*, 5–23. [CrossRef]
65. Manacorda, M.; Tesei, A. Liberation technology: Mobile phones and political mobilization in Africa. *Econometrica* **2020**, *88*, 533–567. [CrossRef]
66. Seo, M.; Shin, Y. Dynamic Panels with Threshold Effect and Endogeneity. *J. Econom.* **2016**, *195*, 169–186. [CrossRef]
67. Jiang, T. Mediating and moderating effects in empirical studies of causal inference. *China Ind. Econ.* **2022**, *380*, 100–120. [CrossRef]
68. Qiu, S.; Wang, Z.; Geng, S. How do environmental regulation and foreign investment behavior affect green productivity growth in the industrial sector? An empirical test based on Chinese provincial panel data. *J. Environ. Manag.* **2021**, *287*, 112282. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Journal of
Theoretical and Applied
Electronic Commerce Research

MDPI

Article

Explaining Policyholders' Chatbot Acceptance with an Unified Technology Acceptance and Use of Technology-Based Model

Jorge de Andrés-Sánchez [1,*] and Jaume Gené-Albesa [2]

[1] Social and Business Research Group, Department of Business Administration, University Rovira i Virgili, Campus de Bellissens, 43204 Reus, Spain
[2] Department of Business Administration, University Rovira i Virgili, Campus de Bellissens, 43204 Reus, Spain; jaume.gene@urv.cat
* Correspondence: jorge.deandres@urv.cat

Abstract: Conversational robots powered by artificial intelligence (AI) are intensively implemented in the insurance industry. This paper aims to determine the current level of acceptance among consumers regarding the use of conversational robots for interacting with insurers and seeks to identify the factors that influence individuals' behavioral intention to engage with chatbots. To explain behavioral intention, we tested a structural equation model based on the Unified Theory of Acceptance and Use of Technology (UTAUT) model. It was supposed that behavioral intention is influenced by performance expectancy, effort expectancy, social influence, and trust, and by the moderating effect of insurance literacy on performance expectancy and effort expectancy. The study reveals a significant overall rejection of robotic technology among respondents. The technology acceptance model tested demonstrates a strong ability to fit the data, explaining nearly 70% of the variance in behavioral intention. Social influence emerges as the most influential variable in explaining the intention to use conversational robots. Furthermore, effort expectancy and trust significantly impact behavioral intention in a positive manner. For chatbots to gain acceptance as a technology, it is crucial to enhance their usability, establish trust, and increase social acceptance among users.

Keywords: insurtech; fintech; chatbots; Unified Theory of Acceptance and Use of Technology; structural equation modeling; partial least squares

1. Introduction

Fintech, which can be defined as the application of technological advances to improve financial products and services [1], has experienced explosive development in recent years because of the impact of digital technologies such as artificial intelligence on the financial industry [2]. Thus, because Insurtech is a branch of fintech that focuses on the insurance sector [3], it can be defined as the application of Industry 4.0 technologies to offer new solutions to the challenges that arise in the provision of products and services as well as in the development of processes in the insurance industry [4]. Although Insurtech has experienced sustained growth since the beginning of the 2010s [5,6], the SARS-CoV-2 crisis has strengthened this trend, as it has caused the introduction of many measures to provide promised services to customers who were hindered by mobility limitations [7].

The implementation of Insurtech must be understood comprehensively, given that Industry 4.0 encompasses all levels of the insurers' value chain, including risk and data analysis, sales, fraud investigation, and asset and liability management [4]. In fact, companies that adopt an integrated approach in the implementation of the digital agenda tend to have better business performance [5]. As outlined in [8], it also embraces customer experience, since the use of instruments powered by artificial intelligence allows for an extremely personalized service.

Industry 4.0 techs have been applied intensively in the insurance industry [6]. This work is focused on the use of conversational robots, whose implementation in the insurance sector began in the mid-2010s [9]. They can be defined as software that mimics human conversations using artificial intelligence algorithms whose purpose is to allow the user to find answers to their queries [10]. It is expected that chatbots will play an important role in improving the services offered by the insurance industry, but their effectiveness will depend on customer satisfaction [11]. Therefore, it is relevant to understand the factors that affect the acceptance of chatbots for carrying out procedures related to current policies.

In theory, chatbots have become an effective tool for addressing user queries in an automated way [12], so many companies have employed them to provide assistance to customers [13]. Conversational bots are promising for enhancing consumer support since they can simplify processes, avoid waits, and offer unrestricted availability around the clock. This enables human agents to concentrate on handling nonroutine and intricate matters, thereby leveraging their expertise effectively [14]. In an insurance environment, chatbots can perform tasks such as searching for suitable products for potential customers based on some keywords, providing information to policyholders about current policies and processes, or speeding up and facilitating claims processes [15]. Several studies report that the implantation of a chatbot system in the organization is judged positively by the stockholders, since news in this regard is followed by an enhancement of the value or the organization [13].

Robotics science is experiencing rapid growth, and chatbots developed with artificial intelligence, which are the most advanced conversational agents, are programmed to interact like real human beings and continuously learn from conversations to provide more accurate answers [12]. In fact, there have been chatbot prototypes capable of passing the Turing test since 2014, when the Eugene Goodman prototype did so [16]. However, the technology behind chatbots is not yet fully mature and often presents flaws [17]. Chatbots are not capable of capturing the evolution of a conversation by interpreting the tones and inflections of the customer's voice, lack emotional skills such as empathy, and cannot respond to complex requirements [18]. These limitations may explain the reluctance of many insurance consumers to interact with chatbots [9,19].

This article evaluates attitudes toward the use of chatbots in insured–insurer interactions in Spain through the use of a structured survey answered by more than 250 persons. The analysis is based on the Unified Theory of Acceptance and Use of Technology (UTAUT) by Venkatesh et al. [20], which has been used in a great number of papers to explain the acceptance of a great number of fintechs [21]. This paper also introduces trust into the analysis, which has been proven to be a key factor in explaining the acceptance of robot technology [22] and the adoption of artificial intelligence technology in the Insurtech setting [23]. We have also taken into account the possible moderating effect of insurance literacy on the effects of performance expectancy and effort expectancy on behavioral intention. This work addresses two research questions:

RQ1. *What is the average intention to be assisted by conversational robots to make procedures such as declaring a claim?*

RQ2. *Does UTAUT, along with trust and insurance literacy, provide a statistically significant framework for the analysis of chatbot acceptance for policy management?*

To answer both questions, we used a structured questionnaire that was answered by 252 respondents. Whereas RQ1 can be answered by simply observing the average punctuation of the questions linked to intention to use bots, the second research question was assessed by adjusting a structural equation model that was fitted using the partial least squares technique.

The rest of the article is structured as follows. In the second section, a review of the literature related to our analysis is carried out, which justifies the use of UTAUT to explain attitudes toward the use of chatbots with the addition of trust and insurance literacy. The third section describes the materials and analytical methodology adopted in the study.

Subsequently, the results of the data analysis are presented. Finally, the main findings are discussed, practical implications are outlined, and the limitations of the study are commented on, suggesting future research.

2. A Unified Technology Acceptance and Use of Technology-Based Model to Assess Behavioral Intention of Policyholders to Use Chatbots

2.1. Initial Considerations

This study tests a model of policyholders' chatbot tech acceptance based on the Unified Theory of Acceptance and Use of Technology (UTAUT) [20] that was refined in [24]. It is a synthesis of previous technology acceptance models, such as the Technology Acceptance Model (TAM) and its extensions [25,26], or the Theory of Planned Behavior [27]. UTAUT links performance expectancy (PE), effort expectancy (EE), and social influence (SI) with the behavioral intention (BI) to use a given technology. The fourth original explanatory factor, facilitating conditions, is supposed to directly impact the actual use of the technology and not on behavioral intention [20], which is the assessed output variable in this work. Likewise, the interaction with bots has negligible digital barriers since even a conventional phone is able to use them. In any case, chatbots have fewer impediments to their use than other digital technologies, such as apps [28]. Therefore, this factor is not considered in this analysis. On the other hand, to explain BI, we added trust. The reason for considering trust is that it has been proven to be a key factor in explaining customer acceptance of chatbots [22] and is a cornerstone factor of the insurance industry [29].

In all business sectors, at the time of writing this paper, the introduction of chatbots to provide customer service is not being implemented uniformly and consistently across all companies [13]. This statement also holds true in the insurance sector [5]. For instance, while some companies may make it mandatory for customers to interact with a chatbot initially, with the option to continue with the bot or request assistance from a human operator based on their satisfaction, other companies allow customers to choose their preferred mode of interaction right from the beginning. Therefore, we understand that the meaning of the variable usage may be heterogeneous in the sample, as it could initially be mandatory for some of the survey respondents. In contrast, "behavioral intention" is a homogeneous concept for all respondents and is particularly relevant for measuring the acceptance of technology by customers. In fact, in technology acceptance models such as the UTAUT model [20,24] or the TAM [25,26], perceived usefulness and effort expectancy influence behavioral intention. Furthermore, the majority of the literature reviewed, which is used to formulate our hypotheses, focuses in some cases primarily on explaining behavioral intention rather than current usage, which in many cases may be driven by mandatory requirements rather than behavioral intention.

It should also be noted that two demographic factors, age and gender, which are considered potential moderators in the work [20], are taken into consideration, but without making any concrete hypothesis about the sign of their moderating capability. The negligible presence of digital barriers in the use of bots makes it difficult to construct any hypothesis regarding the effect of age. Regarding the moderating capability of gender, it should stem from social aspects related to the roles assigned to a specific gender in the society within which the study is framed. In our case, it is difficult to identify gender-related issues to make insurance procedures in the population under study and at the time of the study (the 2020s). Thus, even though in an initial stage we will assess the mediating capacity of the two highlighted sociodemographic factors, only those factors that initially demonstrate a statistically significant effect will be definitively considered in the model.

Likewise, we have also considered possible moderating effects of insurance literacy on the effects of performance expectancy and effort expectancy on behavioral intention, which is supported by arguments provided in [30] in a fintech context and [31] in the evaluation of chatbot acceptance by customers of online travel agencies. It should be noted that, in [20,24], experience is considered a moderating variable, particularly in terms of effort expectancy. In these studies, it is assumed that experience pertains to the technology

being used rather than the task itself, which is the regular job. Therefore, it is implied that experience in the task exists within the target population under study. The variable insurance literacy should be considered in another way in which the construct of experience can be manifested. However, in this case, it pertains to the task being performed (managing active policies), as proficiency in this aspect varies among the target population due to it not being a regular task. On the other hand, using the technology does not require any special experience, as theoretically, it is sufficient to *speak* or *chat* [15,18].

In the reviewed literature on technological acceptance in finance and insurance, as well as the use of chatbots for customer service, the mainstream is built upon the foundation of the TAM or UTAUT, which are adapted to the specific technology and its intended use. While a TAM model suggests that behavioral intention is solely influenced by perceived usefulness and effort expectancy, with other relevant factors such as social influence (or subjective norm) impacting behavioral intention mediated by these basic constructs [26], UTAUT postulates that these constructs, including social influence, directly influence the intention to use [20]. Moreover, following UTAUT, this study does not consider a hypothetical indirect impact of effort expectancy on behavioral intention through performance expectancy [20], which TAM does take into account [25]. Consequently, the model examined in our paper, as illustrated in Figure 1, is grounded in the UTAUT framework rather than provided by TAM.

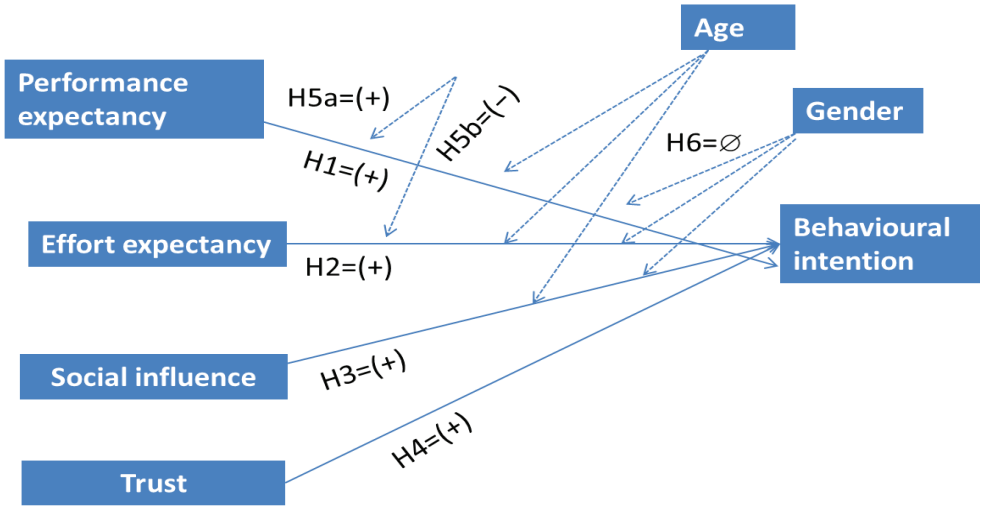


Figure 1. Inner model tested in this paper. Source: Own elaboration with the ground of Venkatesh et al. [20].

2.2. Direct Effects of Performance Expectancy, Effort Expectancy, Social Influence and Trust on Behavioral Intention

The explained outcome in this paper is behavioral intention. Following [20,25], behavioral intention is the main antecedent of usage and can be defined as "persons' readiness to perform a given behavior" that, in our context, consists of interacting with chatbots to implement procedures about existing policies.

Performance expectancy (PE), which is equivalent to so-called perceived usefulness in TAM, can be defined as the degree of perception by potential users that the assessed system will help to improve the attained performance when developing a task [20]. The usefulness of chatbots to make insurance procedures can be justified with several arguments. First, it allows basic demands to be solved more quickly than if exclusively using human assistance [14]. Likewise, bots do not replace other communication channels for interacting with the insurance company but can be understood as an alternative tool to improve service to policyholders [32]. The diversification of communication channels is usually appreciated and fosters customer loyalty [33].

The capability of Insurtech to increase the value of services provided to policyholders and insureds and to enhance production and marketing efficiency [3] must allow the insurer to achieve advantages over its competitors, which could be offering better products and service and/or diminishing prices.

On the other hand, at the time of writing this work, chatbots have several disadvantages that can reduce their usefulness. Generally, chatbots cannot solve complex situations, and therefore, management must be resolved by a human operator [18]. In those cases, the first interaction with the chatbot is perceived as a waste of time and a decrease in its perceived usefulness [17].

Under certain conditions, the utilization of chatbots can have an adverse effect on the perception of the service. When policyholders experience a significant loss, such as the destruction of their home, they anticipate the interlocutor to provide attentive support [34]. Studies have highlighted a notable drawback in employing chatbots for customer service: their inherent absence of empathy [18].

To truly benefit from bot technology, companies must achieve good coordination between data input and processing in the back-end system so that the front-end system can provide reliable responses and appropriate service [9]. Otherwise, the poor quality of front-end system responses results in a poor perception of the usefulness of the interaction [26].

The papers examined reveal a considerable consensus that perceived ease of use (PE) plays a favorable role in shaping individuals' perceptions of technological excellence. This outcome has been observed consistently across various domains, including blockchain applications in finance [35–38], digital banking [39–43], intention to use online applications in the insurance industry [22,44], and acceptance of conversational robot services [10,45–54]. Therefore, the following hypothesis is proposed in this study:

Hypothesis 1 (H1). *The performance expectancy of chatbots positively influences behavioral intention toward their use in customer procedures with the insurance company.*

Following Venkatesh et al. [20], *effort expectancy* (EE) can be defined as the intensity of the belief that to employ a given tech is not difficult, and is essentially the perceived ease of use of TAM [25]. In this study, EE should be understood as the absence of obstacles to communicating and carrying out transactions with the insurance company. Theoretically, chatbots provide some advantages over alternative channels. They are capable of providing assistance 24 h a day, 7 days a week, and are more flexible than humans in terms of when they are available [14]. Moreover, they have fewer barriers to use than other digital technologies, such as apps [28].

On the other hand, there is currently a consensus that chatbot technology is not sufficiently developed to allow for easy interaction in many circumstances. It is very common for chatbots to offer confusing responses to customers, which worsens their perception of usability and usefulness and therefore negatively impacts their acceptance [14]. Some issues that can be highlighted in this regard are technological robotic phobia, which has a significant impact on customers' perception of usability and attitude toward bots [55], and the fact that chatbots do not capture voice tones or the direction of a particular conversation [18].

The drawbacks noted in the above paragraphs lead to bots frequently making errors in real interactions [56]. This problem hinders ease of use [17] and explains why more than half of the interactions with bots are not completed [19]. The literature has also shown the relevance of this variable to explain behavioral intention in contexts similar to those analyzed in this paper, such as blockchain technology [35–38], digital banking [39,42], insurtech [44], and attitudes toward chatbots by customers [22,46–48,52,53,57]. Therefore, the following hypothesis is proposed.

Hypothesis 2 (H2). *The effort expectancy of chatbots positively influences behavioral intention toward their use in customer transactions with the insurer.*

Social influence (SI) is "the degree to which an individual perceives that important others believe they must use a new tech" [20]. In this regard, the opinion of personal

financial advisors and insurance brokers is usually significant in decisions about insurance issues [21]. Although conversational robots have become commonplace in the provision of services to customers, a significant number of consumers still harbor doubt and exhibit hesitance in regard to interacting with them [58]. However, despite this skepticism, there continues to be strong commercial interest in chatbot technology, primarily driven by the recognized advantages it offers. Consequently, businesses actively promote the adoption of conversational bots among their customers [11]. In fact, that tech is commonly employed to deliver initial assistance to both customers and users of public services [18].

It is a widely established fact that the opinion of peers such as friends or family members impacts on the perception and attitude toward new techs because people often act to attain the approval of appreciated persons [20]. This also applies to areas related to financial communication channels [59], new financial instruments such as crowdfunding [60], and the acceptance of conversational bots [47].

As is frequently observed when disruptive technologies are integrated, numerous jobs associated with the insurance industry will become obsolete. Projections indicate that by 2030, the adoption of Industry 4.0 technologies such as chatbots will lead to a drastic reduction of 70% to 80% in the workforce required to cater to customers' needs [61]. This consequence can be perceived as a negative social utility that is common to all Industry 4.0 technologies [62], thus negatively impacting the behavioral attitude toward the tech. Likewise, although many positive benefits can be obtained from techs powered by AI, it is also true that there is an extended social perception that AI implies several dangers, such as job losses. Some examples are that multinational companies may use AI to assemble enormous nonauditable influence and power; the fear that engines powered by AI could become more intelligent than humans, and this jeopardizes humanity; or issues with privacy [63]. Therefore, we propose to test the following hypothesis.

Hypothesis 3 (H3). *Social influence has a positive influence on policyholders' behavioral intention to use chatbots in their procedures with the insurer.*

In empirical studies of technology acceptance, it is common to utilize theoretical frameworks such as the Technology Acceptance Model (TAM) and its extensions [25,26], the Theory of Planned Behavior [27], or UTAUT [20,24], while introducing new variables that are not covered by these theoretical frameworks but are relevant to the analyzed technology and its context. In our case, the relevance of *trust* (TRUST) in the relationships between the insurer and their company is significant, as it encompasses two aspects: the relational and cognitive dimensions [23]. This explains why, within the empirical literature related to our research, it is common to complement the technology acceptance theory used as the theoretical framework with the introduction of trust as a relevant construct. For instance, the acceptance of chatbots by consumers [47] has been studied from within the theory of reasoned action, and [10,46,57] introduce trust in a UTAUT formulation. In studies on the applications of blockchain technology in finance [35,38], TAM is used as the foundation for including trust as an explanatory factor. Within e-banking channels [39,41], trust is evaluated within the TAM framework. Likewise, [64] introduce trust in a UTAUT-based model to evaluate the acceptance of e-payment methods, and [44] also introduces that construct in their TAM model for online insurance evaluation.

Trust should have an impact on policyholders' acceptance of Insurtech in two distinct ways. First, given the unique characteristics of financial and insurance industries, coupled with the fact that the interaction between the company and policyholders is mediated by technology, trust emerges as an exceptionally relevant factor in elucidating customers' attitudes and behavioral intentions [23]. Second, trust grounds financial agreements [64], and this fact is especially true in the insurance industry. Here, both parties involved in the transaction, namely, the insurer and the policyholder/insured, must exhibit trust in each other within an environment characterized by a heightened level of moral hazard and a substantial likelihood of adverse selection. Consequently, trust is essential in insurance market development [29]. Within the realm of insurance, we can define trust on the part of

policyholders in insurance companies as their perception of the ability of these companies to offer secure compensation in the event of a loss and to ensure satisfactory interactions in this regard [29]. This facet of trust is commonly referred to as relational or organizational trust [65].

Similarly, the concept of trust plays a crucial role in understanding the adoption of emerging technologies such as chatbots, particularly in business-to-consumer (B2C) interactions [66]. Individuals often harbor doubts regarding these tools, primarily because of the absence of transparency [47]. In the realm of remote financial services, trust can be defined as customers' confidence in a particular mode of interaction, wherein they believe that companies will be capable of delivering satisfactory service through this channel [39]. This aspect of trust is tied to what is known as cognitive trust [65].

Insurtech such as chatbots have among their objectives to improve the value of products offered to customers [7,15], so it is expected that greater trust in their benefits will imply a more positive attitude toward their adoption and greater perceived usefulness in their services. There are several empirical studies that show that the direct impact of trust on the attitude and intention to use novel technologies is significant in contexts such as blockchain technology [35,38], digital banking [39,41,64], and Insurtech [44]. In the field of chatbot acceptance, we can highlight the works [10,22,46,48,53,57,67]. Therefore, the following hypothesis is proposed:

Hypothesis 4 (H4). *Trust has a positive influence on policyholders' behavioral intention to use chatbots in their procedures with the insurer.*

2.3. The Moderating Effects of Insurance Literacy, Gender, and Age

The significance of experience, which is linked to self-efficacy as a moderator of constructs influencing behavioral intention, has been postulated by widely recognized theories of technology acceptance such as TAM [26] and UTAUT [20,24]. In these works, experience exclusively pertains to the evaluated technology, as it is applied in tasks where it is assumed that the user possesses extensive experience, for instance, in their own work. Conversely, in our case, heterogeneity with respect to experience lies in an aspect where the consumer may lack knowledge (insurance management), but the utilization of technological advancements does not necessitate special technological skills [28]. In such cases, experience and literacy in the relevant domain can also exert influence, either interacting with other factors [68], through moderating effects [31,69], or through mediation [70] on behavioral intention.

Considering that the literature on technology acceptance is more prevalent in finance than in the insurance industry, which is logical due to the former being a broader and more general field, an initial definition of insurance literacy (IL) can be derived by adapting the definition of financial knowledge proposed by Stolper and Walter [71]. Consequently, insurance literacy can be understood as the extent of understanding regarding essential insurance concepts and the ability to apply them in practical situations. In greater detail, [72] identifies insurance literacy as encompassing dimensions such as knowledge and understanding of personal risk exposure, methods for managing risk, functioning of insurance mechanisms, selection of suitable insurance policies, understanding of the rights and obligations of policyholders, and familiarity with the insurance inquiry process. Furthermore, insurance literacy involves the utilization of this knowledge in decision-making processes related to insurance.

In any case, drawing an analogy with the effects of financial literacy on behavioral intention as presented in several papers, it is expected that insurance literacy will have a positive impact on the influence of performance expectancy on behavioral intention. In the field of m-banking, [73] reported the positive influence of financial literacy on intention to use through the mediation of perceived usefulness. Additionally, in the context of mobile banking, [30] argues that individuals with higher levels of financial literacy are more adept at assessing the financial implications of technology adoption, leading to greater performance expectancy, as they perceive clear advantages from its usage. Similarly, in the realm of fintech, [69] found a significant positive impact of literacy on behavioral intention

by moderating perceived usefulness. This moderating effect is also reported in [74] when explaining attitudes toward taking financial risks.

In the specific context of insurance decisions, [75,76] also discovered a positive mediating effect of insurance literacy through factors associated with performance expectations on customers' behavioral intention and attitude toward subscribing to insurance policies.

A similar perspective is presented in a study examining the acceptance of conversational chats in relationships with online travel agencies [31]. This study found that greater familiarity with online travel bookings can positively moderate the impact of expected usefulness on behavioral intention.

Furthermore, experience and knowledge within the field where a new technology is implemented are expected to have a negative moderating role in effort expectancy, as they carry less significance in decision-making for individuals with greater experience [20,26,70]. The moderating influence of financial literacy on effort expectancy when adopting new bank channels has been observed in [69,77]. Based on these findings, the following hypotheses are proposed:

Hypothesis 5a (H5a). *Insurance literacy positively moderates the effect of performance expectancy on behavioral intention*.

Hypothesis 5b (H5b). *Insurance literacy negatively moderates the effect of effort expectancy on behavioral intention.*

Age and gender are considered potential moderators in [20]. This fact motivates us to take the first step in testing the hypothesis presented in Figure 1 by examining their moderating capabilities. In [20], it is justified that gender can serve as a moderating variable for PE, EE, and SI, as men are typically more task-oriented and less adverse to risk compared to women. Conversely, women tend to place greater value on the new system's potential for enhancing ease to use and are more sensitive to subjective norms. Additionally, [20] justifies age as a moderating variable for PE, EE, and SI based on the observation that younger individuals tend to be more responsive to external rewards. Older users may have a reduced ability to interact with new systems while devoting attention to job-related responsibilities. Furthermore, they often possess a stronger sense of belonging to a social group.

However, we believe that the potential mediating role of sociodemographic variables depends on various factors, such as the specific technology being evaluated, the context in which it will be used, and the society in which it is tested. In the case of interacting with conversational bots, where there are no digital impediments compared to apps [28], constructing hypotheses regarding the moderating effect of age becomes challenging, as our context is not directly linked to job tasks.

Regarding the moderating effect of gender, [20,24] emphasize that it should arise from social aspects and the specific roles assigned to a particular gender within the societal framework of the study. In our particular case, it is difficult to identify gender-related issues within the population under study in the year 2023.

Only a limited number of studies have examined the moderating capability of age and gender on performance expectancy (PE), effort expectancy (EE), and social influence (SI) [36,42,52]. In [36], both variables are considered, but the significance of their moderating capacity is not reported. Conversely, in [42], the inclusion of the gender variable is motivated by the potential relevance of gender roles stipulated by religious factors in Kenya, and indeed, this moderation is found to be significant in that study. In [52], the potential moderating capacity of the gender variable is explored in relation to an experiment on the perception of gendered anthropomorphic design cues moderating behavioral intention toward chatbots. However, the results presented in this paper are inconclusive.

Regarding the mediating effect of sociodemographic variables on effort expectancy, there are also very few studies in the reviewed literature. While [36] does not report the results, [42] indicates the significance of gender. However, studies such as [52] and [57] either fail to reach conclusive results or reject the significance of the gender variable. Similar

patterns are observed with regard to social influence. The limited number of studies that have examined the mediating effect of age and gender [42,52] have reported contradictory findings. Thus, we propose the following hypothesis:

Hypothesis 6 (H6). *There is no significant moderating effect of gender and age on the effect of performance expectancy, effort expectancy, and social influence on behavioral intention.*

3. Materials and Methods

3.1. Materials

The survey used a structured questionnaire that was written and provided in Spanish. The English version is shown in Table 1. It began with the following introductory text: "We require your opinion on the management of your current policies regarding contacting the insurer and using automated systems such as voice and text robots instead of a human operator. For example, consider a common procedure such as declaring a claim". The questions were answered on an 11-point Likert scale that ranged from 0 (complete disagreement) to 10 (complete agreement). While it is recognized that the original UTAUT employed a seven-point Likert scale and that subsequent reviewed papers have utilized scales ranging from five to seven grades, current mainstream psychometric scales tend to follow the five–seven-point range [78]. However, a literature review on scales [78] states that scales with five or seven points do not adequately capture the full range of human capability to discern nuances. Typically, people tend to avoid using extreme values when responding, resulting in only one (or two) viable option(s) to express agreement or disagreement for many respondents when using a five (seven)-point scale. Additionally, scores on lower-point scales tend to be higher than those on scales with a greater number of grades. These considerations have led several authors to suggest that an 11-level scale is more suitable than scales with fewer points. Moreover, an 11-point scale allows for an "indifferent" response, and respondents can provide answers in a more intuitive manner, given the prevalence of decimal numerical systems in everyday situations [78]. It should be noted that the implementation of the UTAUT has also utilized 11-point scales in various studies.

For instance, [79] applied the UTAUT theoretical framework to explain the acceptance of cyborg techs, [80] did so in the context of sport technology, and [68] evaluated the behavioral intention to invest in cryptocurrencies. Alesanco-Llorente et al. [81], adopting the cognitive–affective–normative (CAN) theory of technology, employed an 11-point Likert scale to assess behavioral intention toward mobile augmented reality techs, while [82] did the same in the field of wine innovations.

The survey was provided first to six practitioners of the Spanish insurance market who were members of the International Association for Insurance Law (Spanish section). After obtaining their feedback and taking it into account in the new version of the questionnaire, it was distributed to another 12 respondents who were not affiliated with the financial and insurance sector. After considering these latest comments, we established the final questionnaire. The comments from those 18 volunteers improved the readability of the questionnaire but did not imply substantial changes and, moreover, allowed us to make a first test of the measurement scales.

The next step consisted of sharing the survey through social networks (LinkedIn, Telegram, moderated mailing lists, etc.). The survey was completed online from 20 December 2022 to 12 March 2023. Since we were looking for opinions from consumers who were truly informed, we only accepted responses from people who were actually holders of at least two policies. The total number of responses obtained was 252. However, responses where there were unanswered items were not used. Therefore, the final number of responses analyzed in this paper was 226. It allowed a margin of error of $\pm 5\%$ to $\pm 7.5\%$, which, given that the reference population is large (>5000), we judged to be correct.

Table 1. Constructs and items.

Item	Foundation
Behavioral intention BI1. I agree in interacting with a bot to make procedures with my insurer. BI2. I believe that I will employ conversational bots to interact with the insurer regarding my policies. BI3. I will opt to manage existing policies with conversational robots.	Grounded in [20,25]
Performance expectancy PE1. Chatbots make procedures with the insurance company easier. PE2. Chatbots allow for a faster resolution of issues with my policies. PE3. Chatbots allow for making procedures with the insurance company with less effort.	Grounded in [20,24] and un used on [48] in the assessment of AI acceptance
Effort expectancy EE1. Using chatbots to communicate with the insurer is easy. EE2. Managing claims and other procedures with the insurer through chatbots is clear and understandable. EE3. The help of chatbots in managing policies and claims is accessible and less prone to errors. EE4. It is easy to use the communication channels of the insurer smoothly through chatbots.	Based on [20,48,59]
Social influence SI1. Persons who are important for me think that chatbots makes easier insurance procedures. SI2. The people who influence me feel that if there is possibility to choose a channel, better interact with bots. SI3. Persons whose opinions I value feel that making insurance procedures with bots is a step forward.	Grounded in [20] and used by [48,59]
Trust TRUST1. I feel that conversational bots are reliable. TRUST2. The use of chatbots enable the insurance company to fulfil its commitments and obligations. TRUST3. The use of chatbots to interact with the insurer considers the interests of policyholders.	Based on [40] that was grounded in [83]. Used in [57] within a chatbot context.
Insurance literacy IL1. I have a good level of knowledge about insurance matters. IL2. I have a high ability to apply my knowledge about insurance in practice.	Based on [84]

Table 2 shows the profile of the people who responded to the survey. It can be observed that a very high proportion of them are over 40 years old. It can also be seen that the distribution by gender is balanced. Additionally, a percentage close to 50% claim to have more than four policies. Most respondents (approximately 90%) reported being at least college graduates. This high proportion could be attributed to the fact that a significant portion of responses come from LinkedIn, which is composed of a very high proportion of professionals with a university degree. Regarding the most common income profile, it can be considered high (≥EUR 3000, 35%) or medium (between EUR 1750 and EUR 3000), which is consistent with the fact that we were seeking responses from decision-makers with at least two insurance policies.

Table 2. Sociodemographic profile of the respondents.

Gender	**Age**
Male: 52.03%	≤40 years: 14.375
Women: 45.53%	>40 years and <55 years: 53.89%
Other/NA: 2.44%	>55 years: 29.94%
	NA: 1.80%
Academic degree	**Income**
At least graduate: 88.62%	≥EUR 3000: 34.55%
Undergraduate: 11.38%	≥EUR 1750 and <EUR 3000: 35.77%
	<EUR 1750: 29.67%
Number of policies	
>4 contracts: 52.03%	
≥2 and <4: 47.97%	

3.2. Data Analysis

RQ1, *what is average intention to be assisted by conversational robots to make procedures such as declaring a claim*?, was answered by comparing the central tendency measures of the responses, mean and median, of the items of behavioral intention with 5, which can be considered the neutral value on an eleven-point Likert scale. This analysis was implemented using Student's t and Wilcoxon tests.

RQ2: *Does UTAUT, along with trust and insurance literacy, provide a statistically significant framework for the analysis of chatbot acceptance for policy management*? was answered by fitting a structural equation model that was adjusted with partial least squares (PLS-SEM) to the model proposed in Figure 1 and developed in Section 2. One positive aspect of using PLS-SEM is that it can fit complex models without the need for explicit assumptions about the data distribution. Additionally, it does not require a high number of observations [85].

In the initial stage, the model presented in Figure 1 was adjusted, also taking into account the hypothetical mediating effect of age and gender as postulated in the original formulation of UTAUT [20]. This allows the testing of H6. Subsequently, we re-adjusted the model developed in Section 2, considering only the age and gender moderation effects that were found to be statistically significant in the first step. In this regard, age was defined as a dichotomous variable, taking a value of 1 if the individual was not older than 50 years, and 0 otherwise. Gender was also defined as a dichotomous variable, taking a value of 1 if the respondent identified as female and 0 otherwise.

Note that the construct that absorbs the highest number of direct interactions (represented by the arrows) is behavioral intention. It absorbs four direct interactions (from perceived usefulness, effort expectancy, social influence, and trust), along with the additional two interactions corresponding to the mediation of insurance literacy, resulting in a total of six interactions and six more mediation effects that come from age and gender. Therefore, applying the 10 times rule, we can determine that a minimum of 120 observations is required to implement the proposed model using PLS-SEM if all the possible interactions are considered that are reduced to 60 whether effects from age and gender are not under analysis [86].

Additionally, if we consider the inverse R squared method and set a minimum R^2 value of at least 10% for the constructs being predicted (in this case, only behavioral intention), then following [86], for a significance level of 5% and a power of 80%, our sample size also fulfills the criterion, because N > 203.

To fit the model in Figure 1, we used SmartPLS4.0 software following a standard procedure. First, we evaluated the measurement model by verifying the internal consistency of the scales and their discriminant validity. To check internal consistencies, we used the usual measures: Cronbach's alpha, composite reliability (CR), Dijkstra and Henseler's ρ_A, [87] which should all be above 0.7, and the average extracted variance (AVE), which should be greater than 0.5. Furthermore, the factor extraction of the items should be above 0.7. The discriminant validity of the constructs was assessed with Fornell–Larker's criterion [88] and the value of heterotrait–monotrait ratios (HTMT), which follows [89], must be <0.9 [89].

The estimation of path coefficients β is performed using the bootstrapping technique with 5000 subsamples and the studentized method. The p values allowed us to test the hypotheses we proposed in the second section. The overall quality of the fit is quantified with the coefficient of determination R^2, which is complemented by the standardized root mean square residual (SRMR) and normed fit index (NFI). The predictive capability of the model was stated with the Stone–Geisser Q^2 coefficient and, following [90], by using the cross-validated predictive ability test (CVPAT).

4. Results

From the results displayed in Table 3, we can deduce that the measurement scales of BI, PE, EE, SI, TRUST, and IL are reliable because Cronbach's alpha, CR, and $\rho_A > 0.7$, VME > 0.5. Moreover, the factor loadings for all items were >0.7. From Table 4, we can con-

clude that Fornell and Larker's criteria [88] indicate that the constructs have discriminant capability because the correlations of constructs are always below the square root of their AVE. The HTMT ratio criterion also suggests the existence of general discriminant power by the constructs proposed. The exception is the pair PE and TRUST, with a HTMT, despite being <1, of >0.90.

Table 3. Descriptive statistics and analysis of internal consistency of scales.

Items	Mean	Median	Std. Deviation	Factor Loading	Cronbach's α	CR	ρ_A	AVE
BI					0.891	0.895	0.932	0.822
BI1	1.27	0	1.87	0.921				
BI2	2.24	1	2.70	0.861				
BI3	1.38	0	2.06	0.935				
PE					0.918	0.921	0.948	0.859
PE1	2.71	2	2.66	0.929				
PE2	2.57	2	2.58	0.925				
PE3	2.46	2	2.61	0.926				
EE					0.928	0.932	0.949	0.823
EE1	2.88	2.5	2.82	0.869				
EE2	2.69	2	2.61	0.942				
EE3	2.16	2	2.27	0.906				
EE4	2.64	2	2.64	0.910				
SI					0.927	0.929	0.953	0.872
SI1	1.75	1	1.94	0.922				
SI2	1.61	1	2.05	0.953				
SI3	2.03	1	2.15	0.927				
TRUST					0.83	0.865	0.897	0.745
TRUST1	2.07	1	2.50	0.912				
TRUST2	3.46	3	3.04	0.836				
TRUST3	2.08	2	2.18	0.839				
IL					0.969	0.988	0.985	0.971
IL1	5.3	5	2.78	0.988				
IL2	5.05	5	2.88	0.987				

Note: In the three questions regarding behavioral intention, the hypothesis that the mean and median are 5 ($p < 0.01$) has been rejected.

Table 4. Results of discriminant validity with Fornell and Larker (1981) and HTMT ratios by Henseler et al. (2015) criteria.

	BI	PE	EE	SI	TRUST	IL
BI	**0.906**	*0.769*	*0.791*	*0.782*	*0.836*	*0.168*
PE	0.724	**0.909**	*0.838*	*0.685*	*0.948*	*0.194*
EE	0.723	0.799	**0.907**	*0.75*	*0.888*	*0.226*
SI	0.713	0.701	0.641	**0.934**	*0.786*	*0.054*
TRUST	0.734	0.841	0.794	0.69	**0.863**	*0.268*
IL	0.158	0.183	0.215	0.045	0.246	**0.985**

Note: Principal diagonal is the squared root of AVE. Below the principal diagonal are correlations between variables, and above the principal diagonal are HTMT ratios.

Figure 2 shows that the fit of BI reaches an R^2 = 65.6%, which, following [85], can be considered substantial. Note that for these values of the determination coefficients, and under the minimum R^2 criterion, a sample of less than 33 responses would yield significant results [86]. The values of SRMR = 0.054, and thus <0.08, and NFI = 0.85, (i.e., >0.8) also suggest that the model has good adherence to the sample [91].

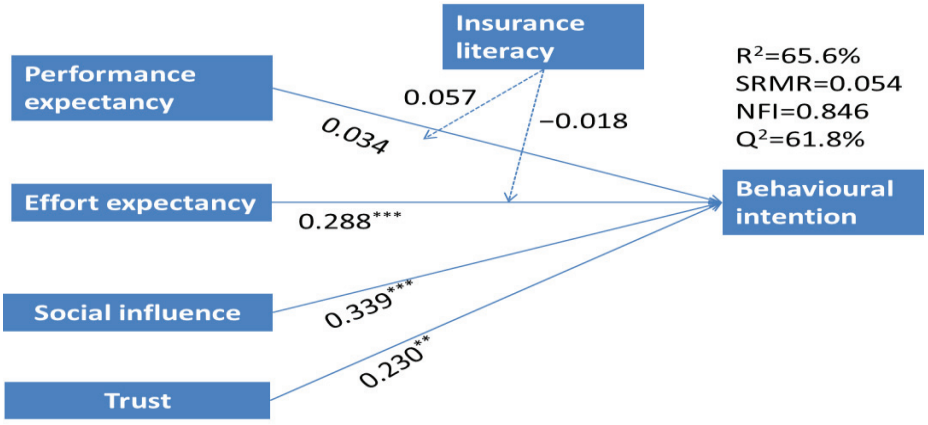


Figure 2. Results of estimating the model described in Section 2 without the moderating effects of age and gender. Note: "***" stands for a statistical significance of $p < 0.01$ and "**" for $p < 0.05$.

Table 5 displays the results of the model presented in Figure 1, considering the potential mediating effects of age and gender on performance expectancy, easiness expectancy, and social influence. It can be observed that, while EE, SI, and TRUST exhibit a significant impact on behavioral intention ($p < 0.05$), this does not hold true for the influence of PE. Furthermore, Table 5 does not present any statistically significant moderation effects of the sociodemographic variables, which are the main focus of analysis in this initial analysis, and so, H6 is accepted.

Table 5. Testing the moderating effects of gender and age over performance expectancy, easiness expectancy, and social influence in the model proposed in Section 2.

Direct effects	β	**Student's t**
PE→BI	−0.036	0.223 (0.823)
EE→BI	0.401	2.556 (0.011)
SI→BI	0.222	1.982 (0.047)
TRUST→BI	0.267	2.749 (0.006)
Moderating effects	β	**Student's t**
PE × IL→BI	0.042	0.457 (0.647)
EE × IL→BI	−0.001	0.007 (0.995)
PE × gender→BI	−0.065	0.391 (0.696)
EE × gender→BI	−0.022	0.126 (0.901)
SI × gender→BI	0.102	0.629 (0.529)
PE × age→BI	0.159	0.979 (0.328)
EE × age→BI	−0.187	1.121 (0.262)
SI × age→BI	0.074	0.498 (0.618)

Note: The *p*-value of the coefficient is in parenthesis.

Figure 2 and Table 6 show the values of β and their level of statistical significance in the final model adopted in this paper, which leads to accepting hypotheses H2, H3, ($p < 0.01$); and H4 ($p < 0.05$). However, unexpectedly, although the sign of the effect of performance expectancy on behavioral intention is as expected, it was not statistically significant, so hypothesis H1 was rejected. We have also found that the sign of the mediation of insurance literacy on PE and EE was as we expected (positive in the first case and negative in the second) but that mediation was not significant, i.e., H5a and H5b were rejected.

Table 6. Path coefficients and results of testing the hypothesis of Section 2.

Effect	β	Student's t	Decision on the Hypothesis
PE→BI	0.034	0.407	H1: Reject
EE→BI	0.288	3.586 ***	H2: Accept
SI→BI	0.339	4.764 ***	H3: Accept
TRUST→BI	0.23	2.441 **	H4: Accept
PE × IL→BI	0.057	0.635	H5a: Reject
EE × IL→BI	−0.018	0.193	H5b: Reject

Note: *** and ** Stands for statistical significance at 1% and 5% probability level, respectively.

Figure 2 and Table 7 show that the structural equation model estimated with partial least squares has a predictive capacity that is at least acceptable. We found that $Q^2 > 0$; therefore, the model provides significant predictions [85]. The use of CVPAT shows that PLS-SEM provides good out-of-sample predictions, as they are more accurate than the benchmark of taking the average of behavioral intention (average loss difference = −2.516, $p < 0.01$). However, we must also acknowledge that the predictions of the presented model do not surpass those of the alternative linear model, so from the perspective of CVPAT, the predictive capacity of our model is not excellent.

Table 7. Predictive capability of the PLS-SEM proposed in Section 2.

		Predictive Measures		Model vs. Average		Model vs. Linear	
	Q^2	RMSE	MAE	ALD	*p* Value	ALD	*p* Value
BI	0.618	0.627	0.456	−2.477	<0.01	0.102	0.463
Overall model				−2.477	<0.01	0.102	0.463

RMSE = root mean square error, MAE = mean absolute error and ALD = average loss difference.

5. Discussion and Implications

5.1. Discussion

This article explains the behavioral intention (BI) of policyholders in managing their existing policies with the help of the UTAUT model [20], along with the consideration of the variable trust, which theoretically should have a significant impact on BI, both due to the nature of the insurance business and the relevance that this factor has in the acceptance of chatbots by consumers. The responses obtained denote a widespread resistance to the use of chatbots, and the proposed model allows for a good explanation of the intention to use since it can explain almost 70% of the variance of the sample. Although the utility of the UTAUT model in the adoption of Fintech has been widely demonstrated, empirical analyses in the Insurtech environment are not as frequent.

Regarding RQ1, *what is average intention to be assisted by conversational robots to make procedures such as declaring a claim*?, a clear attitude of rejection toward interaction with chatbots was observed. This finding is consistent with the literature review conducted in [58] and the results reported in the German insurance market [9] and the Canadian market [19].

Regarding RQ2, *does UTAUT, along with trust and insurance literacy, provide a statistically significant framework for the analysis of chatbot acceptance for policy management*?, we can note that the model proposed in the second section satisfactorily explains the intention to use, as shown by the usual indicators such as the coefficient of determination, standardized root mean square residual, and normed fit index. It also presents relevant predictive capacity, since the value $Q^2 > 0$ and the cross-validated predictive ability test reveal that the proposed model has a predictive capacity superior to the benchmark of taking the average of the BI items. However, it does not make significantly better predictions than the linear benchmark.

Performance expectancy (PE), as indicated by the literature review conducted, is usually the most relevant variable for explaining the adoption of new technologies. Our results show that, although the impact of PE on BI is positive, it does not have statistical significance. Although uncommon, it should be noted that this result has been previously obtained in fields such as m-banking [92], crowdfunding for funding sustainable technology [93], and the adoption of car telematic devices in an insurance setting [94]. Likewise, in the context of chatbot acceptance, [57,67] also do not report a significant direct impact of PE on behavioral intention.

The obtained results show that effort expectancy (EE) has a significant and positive impact on behavioral intention, which is consistent with the revised literature. Along these lines, we found studies related to blockchain applications in finance [35–38], digital banking [39,42], digital insurance intermediation [44], and chatbot adoption by consumers [22,46–48,52,53,57].

The variable with the greatest influence on behavioral intention is undoubtedly social influence (SI). This result is consistent with the studies reviewed in the context of blockchain [35,36], consumers' acceptance of conversational robots [10,44,46–48,51,53,55], m-banking [40,64,92], and innovations related to insurance [21,44].

We have verified that trust also has a significant and positive impact on behavioral intention. This was the expected result given the confluence of the intrinsic nature of the insurance business, which is based on trust [21], and the relevance of this construct in the acceptance of robotic technologies [22], which make this factor highly relevant in Insurtechs powered by AI [23]. Our results coincide with [10,22,46,48,57] in the acceptance of conversational bots, with [35,38] in the field of blockchain technology, with [35,38] in the scope of digital banking, and with [44] in the acceptance of new techs in insurance settings.

It is noteworthy that, on the one hand, as argued in Section 2, the moderating effects of gender and age are not significant, contrary to the postulates of the original UTAUT framework [20]. On the other hand, although the sign of the moderation effect of insurance literacy on the link of performance expectancy and effort expectancy with behavioral intention is as we expected, that finding has no statistical significance. This analytical result contradicts the arguments of H5a and H5b.

The responses, on average, report low ratings both in aspects that potentially influence the formation of behavioral intention (usefulness perception) and in those that have shown statistical significance in our sample (effort expectancy, social influence, and trust). We have observed that the most relevant variable for explaining behavioral intention is social influence. This finding could indicate that in 2023, AI-based technologies are viewed with caution due to the potential problems they can bring, which are exacerbated by the lack of regulation [63], and that disruptive technologies often face significant social opposition [62].

The results also indicate that the development of chatbots for managing existing policies, particularly in areas such as claims, is not sufficiently mature. This explains why surveyed policyholders display resistance to be assisted by conversational bots, since they display low ratings for effort expectancy and trust. This result aligns with the findings of [56], who point out that conversational robots can often create more problems than they solve.

The perception of chatbots observed in our sample aligns with [18] in their study on the use of conversational bots in the relationship between Norwegian citizens and public administrations. Chatbots can be useful in providing assistance for simple customer requests, but require human supervision for slightly more complex issues. Therefore, they can be used by the company as a complementary interaction channel that is able to obtain support from insurance customers in very concrete circumstances.

5.2. Theoretical and Practical Implications

From a theoretical standpoint, we have demonstrated that the combination of the UTAUT analytical framework with trust can capture a significant proportion of the behavioral intention to use chatbots. Note that, whereas we have attained an R^2 of approximately

65%, studies in which TAM and UTAUT are tested in the original formulation [20,25] attain a coefficient of determination that is rarely above 60%. Moreover, the fact that effort expectancy, social influence, and trust are key variables and that insurance culture is not a relevant factor in the behavioral intention to interact with chatbots provides valuable information for firms about the main factors to take into account when implementing conversational chatbots in policyholder services.

We have found that the sociodemographic variables age and gender are not significant in the acceptance of chatbots for assisting policyholders. This finding supports the mainstream literature in the field of technological acceptance in financial and insurance contexts, as well as the use of conversational robots for consumer assistance, where these variables are often not considered in the analysis.

The evolution of technology continues to reshape the insurance industry, with one key development being the emergence of chatbots [3,6]. While chatbots promise efficiency and lower operational costs [15], enhancing the value of insurers in financial markets [13], there exists an underlying reluctance to adopt them [19], as demonstrated in this paper. Our research illustrates that customers still prefer human interaction, considering it more reliable and personalized. Their hesitance to engage with chatbots reflects a broader distrust in automated customer service, partly due to the current limitations of chatbot technology. These bots may lack the capacity to fully understand complex inquiries or complaints, resulting in suboptimal service. This highlights the importance of effort expectancy in understanding the intention to use chatbots in interactions with the insurer. It is important to note that effort expectancy should not be interpreted as a complication in implementing the interaction channel but rather as an assessment of the communication drawbacks.

However, it is essential not to disregard the practical benefits of chatbots. As technology advances, these AI-based entities will become more sophisticated, potentially surpassing human capabilities in terms of speed, accuracy, and availability. Integrating chatbots can also lead to significant cost savings for companies, which could be redirected to enhance customer experiences. This ambivalence necessitates a delicate balance. Companies should prioritize maintaining high customer satisfaction levels in the short term, which may require human-led customer services. A challenge for the insurance industry is to unequivocally demonstrate that the advantages of using chatbots in policyholders' service benefit not only firms but also policyholders and insurers. Thus, the relevance of organizational trust, i.e., the belief that firms' actions are intended to benefit the customer, aligns with this perspective.

Furthermore, it is important to acknowledge that, as consumers encounter chatbots across various services, not just in insurance, their comfort level with this technology is likely to increase. Familiarity can help overcome resistance and improve user acceptance, facilitating a smoother transition toward automated customer service in the insurance sector. Currently, chatbots powered by artificial intelligence are a rising disruptive force that presents several ethical and social concerns in the short term, such as job losses [62], privacy issues, and challenges related to transparency and interpretability [95]. The general social resistance toward emerging disruptive technologies [62], including chatbots, is evident from the high statistical significance of social influence in our empirical application.

This phase of reluctance should be considered an opportunity for the insurance sector to refine their chatbot technology. To achieve this, a closer examination of the customer experience is necessary. By understanding specific customer needs, frustrations, and expectations, companies can tailor their chatbot algorithms to enhance user engagement, increase reliability, and deliver personalized solutions. This approach will make the technology more user-friendly and build trust in the capabilities of chatbots.

In conclusion, the integration of chatbots into the insurance sector is not a matter of "if" but rather "when" and "how". To navigate this transition smoothly, an iterative approach combining technological advancement with the preservation of human elements of customer service may prove most effective. Studies such as the present one can assist

the insurance sector in leveraging the advantages of chatbots without compromising customer satisfaction.

6. Conclusions, Limitations, and Future Research Directions

This study analyzed the perception of policyholders regarding their interaction with conversational bots when performing procedures related to their in-force policies using voice robots. We observed a widespread reluctance toward interacting with voice robots in this context. To establish the drivers of behavioral intention to use chatbots for procedures related to in-force policies, we used the theoretical framework provided by the well-known technology acceptance model UTAUT [20,24]. Specifically, we evaluated the influence of performance expectancy, effort expectancy, and social influence on behavioral intention. Additionally, we included trust and insurance literacy as variables to evaluate, which are not part of the original four variables of the UTAUT. The results show that our model is useful for explaining the behavioral intention to use bots.

We are aware of the limitations of the empirical work presented. This study was conducted in a specific country, Spain, and most responses were obtained from social networks such as LinkedIn. The users of these networks are often individuals with university degrees and a professional background that typically ranges from middle management to executive positions. Thus, the level of education and economic position may bias our conclusions about the behavioral intention to use chatbots. Therefore, it is advisable to exercise caution when extrapolating our findings to policyholders/insureds from cultures that are not close to the Spanish culture or to individuals with professional and educational profiles different from those of the sample used. To obtain broader conclusions, it would be necessary to expand the number of countries represented and the socioeconomic profiles of the surveyed individuals.

Another limitation of this paper is its exclusive reliance on correlational quantitative methods. However, in order to achieve a deeper understanding of the behavioral intention to use chatbot services in financial settings, it would be beneficial to enhance the study with a configurational approach that can identify profiles associated with acceptance and resistance towards chatbot assistance [96]. Additionally, conducting a qualitative analysis of the arguments provided by the surveyed individuals would also be valuable [97]. Undoubtedly, in this regard, UTAUT offers a robust theoretical framework to support such analyses [24].

The conclusions extracted in this study cannot be generalized to the medium and long term because it is grounded on a cross-sectional survey. Industry 4.0 technology and artificial intelligence are fields in which advances are extremely rapid, and attitudes and intention to use a given technology are highly volatile. A more comprehensive perspective will require similar analyses at different stages of chatbot technology development.

Since late 2022 and early 2023, chat systems based on AI technologies such as natural language processing, exemplified by ChatGPT, have experienced exponential development. Prominent figures in the field of information technology, such as Bill Gates, consider this to be one of the most disruptive technological advancements since the 1980s [96].

Undoubtedly, these technologies offer advantages such as their ability to simulate human conversation, user-friendly interfaces, and the capacity to comprehend the contextual nuances of a conversation with a user based on previous interactions [98,99]. However, they also present disadvantages, including the lack of continuous knowledge updating, as ChatGPT is not connected to databases, and the potential for biased or inaccurate responses to the extent that the training databases contain such biases [99].

The characteristics of this technology suggest that its implementation could enhance the consumer experience [98], which, in the insurance industry, could have implications for both the underwriting phase (e.g., assisting in selecting the most suitable coverage based on consumer profiles) and the management of existing policies, which is the focus of this study. Nevertheless, it is essential to consider the drawbacks resulting from the aforementioned issues. Therefore, a natural extension of this research could involve analyzing the potential

acceptance of ChatGPT-like technologies for decision-making support and management in the insurance domain from the perspectives of both insurers and policyholders.

Author Contributions: Conceptualization: J.d.A.-S. and J.G.-A.; methodology: J.d.A.-S.; validation: J.G.-A.; formal analysis: J.d.A.-S.; investigation: J.d.A.-S. and J.G.-A.; resources: J.d.A.-S.; data curation: J.G.-A.; writing—original draft preparation: J.d.A.-S.; writing—review and editing: J.G.-A.; visualization: J.G.-A.; supervision: J.G.-A.; project administration: J.G.-A.; funding acquisition: J.d.A.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research has benefited from the Research Project of the Spanish Ministry of Science and Technology "Sostenibilidad, digitalización e innovación: nuevos retos en el derecho del seguro" (PID2020-117169GB-I00).

Institutional Review Board Statement: (1) All participants received detailed written information about the study and procedure; (2) no data directly or indirectly related to the health of the subjects were collected, and therefore the Declaration of Helsinki was not mentioned when informing the subjects; (3) the anonymity of the collected data was ensured at all times; (4) the research received a favorable evaluation from the Ethics Committee of the researchers' institution (CEIPSA-2022-PR-0005).

Informed Consent Statement: All respondents gave permission for the processing of their responses.

Data Availability Statement: Data available by requiring it to the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Barbu, C.M.; Florea, D.L.; Dabija, D.-C.; Barbu, M.C.R. Customer Experience in Fintech. *J. Theor. Appl. Electron. Commer. Res.* **2021**, *16*, 1415–1433. [CrossRef]
2. Tello-Gamarra, J.; Campos-Teixeira, D.; Longaray, A.A.; Reis, J.; Hernani-Merino, M. Fintechs and Institutions: A Systematic Literature Review and Future Research Agenda. *J. Theor. Appl. Electron. Commer. Res.* **2022**, *17*, 722–750. [CrossRef]
3. Stoeckli, E.; Dremel, C.; Uebernickel, F. Exploring characteristics and transformational capabilities of InsurTech innovations to understand insurance value creation in a digital world. *Electron. Mark.* **2018**, *28*, 287–305. [CrossRef]
4. Yan, T.C.; Schulte, P.; Chuen, D.L.K. InsurTech and FinTech: Banking and Insurance Enablement. In *Handbook of Blockchain, Digital Finance, and Inclusion*; Academic Press: Cambridge, MA, USA, 2018; Volume 1, pp. 249–281. [CrossRef]
5. Bohnert, A.; Fritzsche, A.; Gregor, S. Digital agendas in the insurance industry: The importance of comprehensive approaches. *Geneva Pap. Risk Insur.-Issues Pract.* **2019**, *44*, 1–19. [CrossRef]
6. Sosa, I.; Montes, Ó. Understanding the InsurTech dynamics in the transformation of the insurance sector. *Risk Manag. Insur. Rev.* **2022**, *25*, 35–68. [CrossRef]
7. Lanfranchi, D.; Grassi, L. Examining insurance companies' use of technology for innovation. *Geneva Pap. Risk Insur.-Issues Pract.* **2022**, *47*, 520–537. [CrossRef]
8. Yıldız, E.; Güngör Şen, C.; Işık, E.E. A Hyper-Personalized Product Recommendation System Focused on Customer Segmentation: An Application in the Fashion Retail Industry. *J. Theor. Appl. Electron. Commer. Res.* **2023**, *18*, 571–596. [CrossRef]
9. Rodríguez-Cardona, D.; Werth, O.; Schönborn, S.; Breitner, M.H. A mixed methods analysis of the adoption and diffusion of Chatbot Technology in the German insurance sector. In Proceedings of the Twenty-Fifth Americas Conference on Information Systems, Cancun, Mexico, 15–17 August 2019.
10. Joshi, H. Perception and Adoption of Customer Service Chatbots among Millennials: An Empirical Validation in the Indian Context. In Proceedings of the 17th International Conference on Web Information Systems and Technologies–WEBIST 2021, online, 26–28 October 2021; pp. 197–208. [CrossRef]
11. Xu, Y.; Zhang, J.; Deng, G. Enhancing customer satisfaction with chatbots: The influence of communication styles and consumer attachment anxiety. *Front. Psychol.* **2022**, *13*, 4266. [CrossRef]
12. Nirala, K.K.; Singh, N.K.; Purani, V.S. A survey on providing customer and public administration based services using AI: Chatbot. *Multimed. Tools Appl.* **2022**, *81*, 22215–22246. [CrossRef]
13. Fotheringham, D.; Wiles, M.A. The effect of implementing chatbot customer service on stock returns: An event study analysis. *J. Acad. Mark. Sci.* **2022**, *51*, 802–822. [CrossRef]
14. DeAndrade, I.M.; Tumelero, C. Increasing customer service efficiency through artificial intelligence chatbot. *Rev. Gestão* **2022**, *29*, 238–251. [CrossRef]
15. Riikkinen, M.; Saarijärvi, H.; Sarlin, P.; Lähteenmäki, I. Using artificial intelligence to create value in insurance. *Int. J. Bank Mark.* **2018**, *36*, 1145–1168. [CrossRef]
16. Warwick, K.; Shah, H. Can machines think? A report on Turing test experiments at the Royal Society. *J. Exp. Theor. Artif. Intell.* **2016**, *28*, 989–1007. [CrossRef]

17. de Sá Siqueira, M.A.; Müller, B.C.N.; Bosse, T. When Do We Accept Mistakes from Chatbots? The Impact of Human-Like Communication on User Experience in Chatbots That Make Mistakes. *Int. J. Hum.–Comput. Interact.* **2023**. [CrossRef]
18. Vassilakopoulou, P.; Haug, A.; Salvesen, L.M.; Pappas, I.O. Developing human/AI interactions for chat-based customer services: Lessons learned from the Norwegian government. *Eur. J. Inf. Syst.* **2023**, *32*, 10–22. [CrossRef]
19. PromTep, S.P.; Arcand, M.; Rajaobelina, L.; Ricard, L. From What Is Promised to What Is Experienced with Intelligent Bots. In *Advances in Information and Communication, Proceedings of the 2021 Future of Information and Communication Conference (FICC), Virtual, 29–30 April 2021*; Arai, K., Ed.; Springer International Publishing: Cham, Switzerland, 2021; Volume 1, pp. 560–565.
20. Venkatesh, V.; Morris, M.G.; Davis, G.B.; Davis, F.D. User acceptance of information technology: Toward a unified view. *MIS Q.* **2003**, *27*, 425–478. [CrossRef]
21. Andrés-Sánchez, J.; González-Vila Puchades, L.; Arias-Oliva, M. Factors influencing policyholders′ acceptance of life settlements: A technology acceptance model. *Geneva Pap. Risk Insur.-Issues Pract.* **2021**. [CrossRef]
22. Mostafa, R.B.; Kasamani, T. Antecedents and consequences of chatbot initial trust. *Eur. J. Mark.* **2022**, *56*, 1748–1771. [CrossRef]
23. Zarifis, A.; Cheng, X. A model of trust in Fintech and trust in Insurtech: How Artificial Intelligence and the context influence it. *J. Behav. Exp. Financ.* **2022**, *36*, 100739. [CrossRef]
24. Venkatesh, V.; Thong, J.Y.L.; Xu, X. Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Q.* **2012**, *36*, 157–178. [CrossRef]
25. Davis, F.D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* **1989**, *13*, 319–340. [CrossRef]
26. Venkatesh, V.; Bala, H. Technology acceptance model 3 and a research agenda on interventions. *Decis. Sci.* **2008**, *39*, 273–315. [CrossRef]
27. Ajzen, I. The theory of planned behavior. *Organ. Behav. Hum. Decis. Proces.* **1991**, *50*, 179–211. [CrossRef]
28. Koetter, F.; Blohm, M.; Drawehn, J.; Kochanowski, M.; Goetzer, J.; Graziotin, D.; Wagner, S. Conversational agents for insurance companies: From theory to practice. In *Agents and Artificial Intelligence: 11th International Conference, ICAART 2019, Prague, Czech Republic, 19–21 February 2019*; Van den Herik, J., Rocha, A.P., Steels, L., Eds.; Revised Selected Papers 11; Springer International Publishing: Cham, Switzerland, 2019; pp. 338–362.
29. Guiso, L. Trust and insurance. *Geneva Pap. Risk Insur.-Issues Pract.* **2021**, *46*, 509–512. [CrossRef]
30. Alalwan, A.A.; Dwivedi, Y.K.; Rana, N.P.; Williams, M.D. Consumer adoption of mobile banking in Jordan: Examining the role of usefulness, ease of use, perceived risk and self-efficacy. *J. Enterp. Inf. Manag.* **2017**, *30*, 522–550. [CrossRef]
31. Zhu, Y.; Zhang, R.; Zou, Y.; Jin, D. Investigating customers' responses to artificial intelligence chatbots in online travel agencies: The moderating role of product familiarity. *J. Hosp. Tour. Technol.* **2023**, *14*, 208–224. [CrossRef]
32. Standaert, W.; Muylle, S. Framework for open insurance strategy: Insights from a European study. *Geneva Pap. Risk Insur.-Issues Pract.* **2022**, *47*, 643–668. [CrossRef]
33. Gené-Albesa, J. Interaction channel choice in a multichannel environment, an empirical study. *Int. J. Bank Mark.* **2007**, *25*, 490–506. [CrossRef]
34. LMI Group. The Psychology of Claims. 2022. Available online: https://lmigroup.io/the-psychology-of-claims/ (accessed on 12 December 2022).
35. Albayati, H.; Kim, S.K.; Rho, J.J. Accepting financial transactions using blockchain technology and cryptocurrency: A customer perspective approach. *Technol. Soc.* **2020**, *62*, 101320. [CrossRef]
36. Nuryyev, G.; Wang, Y.-P.; Achyldurdyyeva, J.; Jaw, B.-S.; Yeh, Y.-S.; Lin, H.-T.; Wu, L.-F. Blockchain Technology Adoption Behavior and Sustainability of the Business in Tourism and Hospitality SMEs: An Empirical Study. *Sustainability* **2020**, *12*, 1256. [CrossRef]
37. Sheel, A.; Nath, V. Blockchain technology adoption in the supply chain (UTAUT2 with risk)–evidence from Indian supply chains. *Int. J. Appl. Manag. Sci.* **2020**, *12*, 324–346. [CrossRef]
38. Palos-Sánchez, P.; Saura, J.R.; Ayestaran, R. An Exploratory Approach to the Adoption Process of Bitcoin by Business Executives. *Mathematics* **2021**, *9*, 355. [CrossRef]
39. Bashir, I.; Madhavaiah, C. Consumer attitude and behavioral intention toward Internet banking adoption in India. *J. Indian Bus. Res.* **2015**, *7*, 67–102. [CrossRef]
40. Farah, M.F.; Hasni, M.J.S.; Abbas, A.K. Mobile-banking adoption: Empirical evidence from the banking sector in Pakistan. *Int. J. Bank Mark.* **2018**, *36*, 1386–1413. [CrossRef]
41. Sánchez-Torres, J.A.; Canada, F.-J.A.; Sandoval, A.V.; Alzate, J.-A.S. E-banking in Colombia: Factors favoring its acceptance, online trust and government support. *Int. J. Bank Mark.* **2018**, *36*, 170–183. [CrossRef]
42. Warsame, M.H.; Ireri, E.M. Moderation effect on mobile microfinance services in Kenya: An extended UTAUT model. *J. Behav. Exp. Financ.* **2018**, *18*, 67–75. [CrossRef]
43. Hussain, M.; Mollik, A.T.; Johns, R.; Rahman, M.S. M-payment adoption for bottom of pyramid segment: An empirical investigation. *Int. J. Bank Mark.* **2019**, *37*, 362–381. [CrossRef]
44. Huang, W.S.; Chang, C.T.; Sia, W.Y. An empirical study on the consumers' willingness to insure online. *Pol. J. Manag. Stud.* **2019**, *20*, 202–212. [CrossRef]
45. Eeuwen, M.V. Mobile Conversational Commerce: Messenger Chatbots as the Next Interface between Businesses and Consumers. Master's Thesis, University of Twente, Enschede, The Netherlands, 2017.

46. Kuberkar, S.; Singhal, T.K. Factors influencing adoption intention of AI powered chatbot for public transport services within a smart city. *Int. J. Emerg. Technol.* **2020**, *11*, 948–958.
47. Brachten, F.; Kissmer, T.; Stieglitz, S. The acceptance of chatbots in an enterprise context-A survey study. *Int. J. Inf. Manag.* **2021**, *60*, 102375. [CrossRef]
48. Gansser, O.A.; Reich, C.S. A new acceptance model for artificial intelligence with extensions to UTAUT2: An empirical study in three segments of application. *Technol. Soc.* **2021**, *65*, 101535. [CrossRef]
49. Melián-González, S.; Gutiérrez-Taño, D.; Bulchand-Gidumal, J. Predicting the intentions to use chatbots for travel and tourism. *Curr. Issues Tour.* **2021**, *24*, 192–210. [CrossRef]
50. Balakrishnan, J.; Abed, S.S.; Jones, P. The role of meta-UTAUT factors, perceived anthropomorphism, perceived intelligence, and social self-efficacy in chatbot-based services? *Technol. Forecast. Soc. Chang.* **2022**, *180*, 121692. [CrossRef]
51. Lee, S.; Oh, J.; Moon, W.-K. Adopting Voice Assistants in Online Shopping: Examining the Role of Social Presence, Performance Risk, and Machine Heuristic. *Int. J. Hum.–Comput. Interact.* **2022**. [CrossRef]
52. Pawlik, V.P. Design Matters! How Visual Gendered Anthropomorphic Design Cues Moderate the Determinants of the Behavioral Intention Toward Using Chatbots. In *Chatbot Research and Design, Proceedings of the 5th International Workshop, CONVERSATIONS 2021, Virtual Event, 23–24 November 2021*; Følstad, A., Araujo, T., Papadopoulos, S., Law, E.L.-C., Luger, E., Goodwin, M., Brandtzaeg, P.B., Eds.; Revised Selected Papers; Springer: Cham, Switzerland, 2022. [CrossRef]
53. Silva, F.A.; Shojaei, A.S.; Barbosa, B. Chatbot-Based Services: A Study on Customers' Reuse Intention. *J. Theor. Appl. Electron. Commer. Res.* **2023**, *18*, 457–474. [CrossRef]
54. Xie, C.; Wang, Y.; Cheng, Y. Does Artificial Intelligence Satisfy You? A Meta-Analysis of User Gratification and User Satisfaction with AI-Powered Chatbots. *Int. J. Hum. Comput. Interact.* **2022**. [CrossRef]
55. Rajaobelina, L.; PromTep, S.P.; Arcand, M.; Ricard, L. Creepiness: Its antecedents and impact on loyalty when interacting with a chatbot. *Psychol. Mark.* **2021**, *38*, 2339–2356. [CrossRef]
56. Xing, X.; Song, M.; Duan, Y.; Mou, J. Effects of different service failure types and recovery strategies on the consumer response mechanism of chatbots. *Technol. Soc.* **2022**, *70*, 102049. [CrossRef]
57. Kasilingam, D.L. Understanding the attitude and intention to use smartphone chatbots for shopping. *Technol. Soc.* **2020**, *62*, 101280. [CrossRef]
58. Van Pinxteren, M.M.; Pluymaekers, M.; Lemmink, J.G. Human-like communication in conversational agents: A literature review research agenda. *J. Serv. Manag.* **2020**, *31*, 203–225. [CrossRef]
59. Makanyeza, C.; Mutambayashata, S. Consumers' acceptance and use of plastic money in Harare, Zimbabwe: Application of the unified theory of acceptance and use of technology 2. *Int. J. Bank Mark.* **2018**, *36*, 379–392. [CrossRef]
60. Fanea-Ivanovici, M.; Baber, H. The role of entrepreneurial intentions, perceived risk and perceived trust in crowdfunding intentions. *Eng. Econ.* **2021**, *32*, 433–445. [CrossRef]
61. Balasubramanian, R.; Libarikian, A.; McElhaney, D. Insurance 2030-the Impact of AI on the Future of Insurance. McKinsey & Company. Consultado en. 2018. Available online: https://www.mckinsey.com/industries/financial-services/our-insights/insurance-2030-the-impact-of-ai-on-the-future-of-insurance#/ (accessed on 12 February 2023).
62. Kovacs, O. The dark corners of industry 4.0–Grounding economic governance 2.0. *Technol. Soc.* **2018**, *55*, 140–145. [CrossRef]
63. Stahl, B.C. (Ed.) Perspectives in Artificial Intelligence. In *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*; Springer Nature: Berlin/Heidelberg, Germany, 2021; pp. 7–17. [CrossRef]
64. Tomić, N.; Kalinić, Z.; Todorović, V. Using the UTAUT model to analyze user intention to accept electronic payment systems in Serbia. *Port. Econ. J.* **2023**, *22*, 251–270. [CrossRef]
65. Glikson, E.; Woolley, A.W. Human trust in artificial intelligence: Review of empirical research. *Acad. Manag. Ann.* **2020**, *14*, 627–660. [CrossRef]
66. Baabdullah, A.M.; Alalwan, A.A.; Algharabat, R.S.; Metri, B.; Rana, N.P. Virtual agents and flow experience: An empirical examination of AI-powered chatbots. *Technol. Forecast. Soc. Chang.* **2022**, *181*, 121772. [CrossRef]
67. De Cicco, R.; Iacobucci, S.; Aquino, A.; Romana Alparone, F.; Palumbo, R. Understanding Users' Acceptance of Chatbots: An Extended TAM Approach. In *Chatbot Research and Design, Proceedings of the 5th International Workshop, CONVERSATIONS 2021, Virtual Event, 23–24 November 2021*; Følstad, A., Araujo, T., Papadopoulos, S., Law, E.L.-C., Luger, E., Goodwin, M., Brandtzaeg, P.B., Eds.; Springer: Cham, Switzerland, 2022; Revised Selected Papers. [CrossRef]
68. Arias-Oliva, M.; de Andrés-Sánchez, J.; Pelegrín-Borondo, J. Fuzzy set qualitative comparative analysis of factors influencing the use of cryptocurrencies in Spanish households. *Mathematics* **2021**, *9*, 324. [CrossRef]
69. Akbar, Y.R.; Zainal, H.; Basriani, A.; Zainal, R. Moderate Effect of Financial Literacy during the COVID-19 Pandemic in Technology Acceptance Model on the Adoption of Online Banking Services. *Bp. Int. Res. Crit. Inst. J.* **2021**, *4*, 11904–11915. [CrossRef]
70. Hsieh, P.J.; Lai, H.M. Exploring people's intentions to use the health passbook in self-management: An extension of the technology acceptance and health behavior theoretical perspectives in health literacy. *Technol. Forecast. Soc. Chang.* **2020**, *161*, 120328. [CrossRef]
71. Stolper, O.A.; Walter, A. Financial literacy financial, advice and financial behavior. *J. Bus. Econ.* **2017**, *87*, 581–643. [CrossRef]
72. Sanjeewa, W.S.; Hongbing, O. Consumers' insurance literacy: Literature review, conceptual definition, and approach for a measurement instrument. *Eur. J. Bus. Manag.* **2019**, *11*, 49–65. [CrossRef]

73. Ullah, S.; Kiani, U.S.; Raza, B.; Mustafa, A. Consumers' Intention to Adopt m-payment/m-banking: The Role of Their Financial Skills and Digital Literacy. *Front. Psychol.* **2022**, *13*, 873708. [CrossRef] [PubMed]
74. Grable, J.E.; Rabbani, A. The Moderating Effect of Financial Knowledge on Financial Risk Tolerance. *J. Risk Financ. Manag.* **2023**, *16*, 137. [CrossRef]
75. Nomi, M.; Sabbir, M.M. Investigating the factors of consumers' purchase intention towards life insurance in Bangladesh: An application of the theory of reasoned action. *Asian Acad. Manag. J.* **2020**, *25*, 135–165. [CrossRef]
76. Weedige, S.S.; Ouyang, H.; Gao, Y.; Liu, Y. Decision making in personal insurance: Impact of insurance literacy. *Sustainability* **2019**, *11*, 6795. [CrossRef]
77. Onay, C.; Aydin, G.; Kohen, S. Overcoming resistance barriers in mobile banking through financial literacy. *Int. J. Mob. Commun.* **2023**, *21*, 341–364. [CrossRef]
78. Bisquerra Alzina, R.; Pérez Escoda, N. Can Likert scales increase in sensitivity? *REIRE* **2015**, *8*, 129–147. [CrossRef]
79. Pelegrin-Borondo, J.; Reinares-Lara, E.; Olarte-Pascual, C. Assessing the acceptance of technological implants (the cyborg): Evidences and challenges. *Comput. Hum. Behav.* **2017**, *70*, 104–112. [CrossRef]
80. Andrés-Sanchez, J.; de Torres-Burgos, F.; Arias-Oliva, M. Why disruptive sport competition technologies are used by amateur athletes? An analysis of Nike Vaporfly shoes. *J. Sport Health Res.* **2023**, *15*, 197–214. [CrossRef]
81. Alesanco-Llorente, M.; Reinares-Lara, E.; Pelegrín-Borondo, J.; Olarte-Pascual, C. Mobile-assisted showrooming behavior and the (r) evolution of retail: The moderating effect of gender on the adoption of mobile augmented reality. *Technol. Forecast. Soc. Chang.* **2023**, *191*, 122514. [CrossRef]
82. Reinares-Lara, E.; Pelegrín-Borondo, J.; Olarte-Pascual, C.; Oruezabala, G. The role of cultural identity in acceptance of wine innovations in wine regions. *Br. Food J.* **2023**, *125*, 869–885. [CrossRef]
83. Morgan, R.M.; Hunt, S.D. The commitment-trust theory of relationship marketing. *J. Mark.* **1994**, *58*, 20–38. [CrossRef]
84. Hastings, J.S.; Madrian, B.C.; Skimmyhorn, B. Financial literacy, financial education, and economic outcomes. *Annu. Rev. Econ.* **2013**, *5*, 347–375. [CrossRef] [PubMed]
85. Hair, J.F.; Risher, J.J.; Sarstedt, M.; Ringle, C.M. When to use and how to report the results of PLS-SEM. *Eur. Bus. Rev.* **2019**, *31*, 2–24. [CrossRef]
86. Kock, N.; Hadaya, P. Minimum sample size estimation in PLS-SEM: The inverse square root and gamma-exponential methods. *Inf. Syst. J.* **2018**, *28*, 227–261. [CrossRef]
87. Dijkstra, T.K.; Henseler, J. Consistent partial least squares path modeling. *MIS Q.* **2015**, *39*, 297–316. Available online: https://www.jstor.org/stable/26628355 (accessed on 20 May 2023). [CrossRef]
88. Fornell, C.; Larcker, D.F. Evaluating structural equation models with unobservable variables and measurement error. *J. Mark. Res.* **1981**, *18*, 39–50. [CrossRef]
89. Henseler, J.; Ringle, C.M.; Sarstedt, M. A new criterion for assessing discriminant validity in variance-based structural equation modeling. *J. Acad. Mark. Sci.* **2015**, *43*, 115–135. [CrossRef]
90. Liengaard, B.D.; Sharma, P.N.; Hult, G.T.M.; Jensen, M.B.; Sarstedt, M.; Hair, J.F.; Ringle, C.M. Prediction: Coveted, yet forsaken? Introducing a cross-validated predictive ability test in partial least squares path modeling. *Decis. Sci.* **2021**, *52*, 362–392. [CrossRef]
91. Sarstedt, M.; Ringle, C.M.; Smith, D.; Reams, R.; Hair, J.F., Jr. Partial least squares structural equation modeling (PLS-SEM): A useful tool for family business researchers. *J. Fam. Bus. Strategy* **2014**, *5*, 105–115. [CrossRef]
92. Mahfuz, M.A.; Khanam, L.; Mutharasu, S.A. The influence of website quality on m-banking services adoption in Bangladesh: Applying the UTAUT2 model using PLS. In Proceedings of the 2016 International Conference on Electrical, Electronics, and Optimization Techniques, Chennai, India, 3–5 March 2016; pp. 2329–2335. [CrossRef]
93. Moon, Y.; Hwang, J. Crowdfunding as an alternative means for funding sustainable appropriate technology: Acceptance determinants of backers. *Sustainability* **2018**, *10*, 1456. [CrossRef]
94. Milanović, N.; Milosavljević, M.; Benković, S.; Starčević, D.; Spasenić, Ž. An Acceptance Approach for Novel Technologies in Car Insurance. *Sustainability* **2020**, *12*, 10331. [CrossRef]
95. Rimban, E. Challenges and Limitations of ChatGPT and Other Large Language Models Challenges. Available online: https://ssrn.com/abstract=4454441 (accessed on 20 May 2023).
96. Andres-Sanchez, J.; Almahameed, A.A.; Arias-Oliva, M.; Pelegrin-Borondo, J. Correlational and Configurational Analysis of Factors Influencing Potential Patients' Attitudes toward Surgical Robots: A Study in the Jordan University Community. *Mathematics* **2022**, *10*, 4319. [CrossRef]
97. Andrés-Sánchez, J.; Gené-Albesa, J. Assessing Attitude and Behavioral Intention toward Chatbots in an Insurance Setting: A Mixed Method Approach. *Int. J. Hum.–Comput. Interact.* **2023**. [CrossRef]
98. Rampton, J. The Advantages and Disadvantages of ChatGPT. 2023. Available online: https://www.entrepreneur.com/growth-strategies/the-advantages-and-disadvantages-of-chatgpt/450268 (accessed on 6 June 2023).
99. Deng, J.; Lin, Y. The Benefits and Challenges of ChatGPT: An Overview. *Front. Comput. Intell. Syst.* **2022**, *2*, 81–83. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

big data and cognitive computing

MDPI

Article

Predicting the Price of Bitcoin Using Sentiment-Enriched Time Series Forecasting

Markus Frohmann [1,2], Manuel Karner [1], Said Khudoyan [1], Robert Wagner [1] and Markus Schedl [1,2,*]

[1] Multimedia Mining and Search Group, Institute of Computational Perception, Johannes Kepler University Linz (JKU), 4040 Linz, Austria; markus.frohmann@gmail.com (M.F.)
[2] Human-Centered AI Group, AI Laboratory, Linz Institute of Technology (LIT), 4040 Linz, Austria
* Correspondence: markus.schedl@jku.at

Abstract: Recently, various methods to predict the future price of financial assets have emerged. One promising approach is to combine the historic price with sentiment scores derived via sentiment analysis techniques. In this article, we focus on predicting the future price of Bitcoin, which is currently the most popular cryptocurrency. More precisely, we propose a hybrid approach, combining time series forecasting and sentiment prediction from microblogs, to predict the intraday price of Bitcoin. Moreover, in addition to standard sentiment analysis methods, we are the first to employ a fine-tuned BERT model for this task. We also introduce a novel weighting scheme in which the weight of the sentiment of each tweet depends on the number of its creator's followers. For evaluation, we consider periods with strongly varying ranges of Bitcoin prices. This enables us to assess the models w.r.t. robustness and generalization to varied market conditions. Our experiments demonstrate that BERT-based sentiment analysis and the proposed weighting scheme improve upon previous methods. Specifically, our hybrid models that use linear regression as the underlying forecasting algorithm perform best in terms of the mean absolute error (MAE of 2.67) and root mean squared error (RMSE of 3.28). However, more complicated models, particularly long short-term memory networks and temporal convolutional networks, tend to have generalization and overfitting issues, resulting in considerably higher MAE and RMSE scores.

Keywords: time series forecasting; sentiment analysis; emotion detection; regression analysis; data mining; social networks; Bitcoin

1. Introduction

Bitcoin (BTC) is a decentralized digital currency that allows for peer-to-peer transactions without the need for a central authority or middleman [1]. It was created in 2008 by an individual or group of individuals going by the alias Satoshi Nakamoto. The basic aim behind Bitcoin's invention was to build a decentralized, trustless transaction system that would eliminate the need for middlemen such as banks. Transactions are recorded on a public ledger known as the blockchain, which employs complicated algorithms to maintain the integrity and security of the system. It employs a proof-of-work consensus process to ensure that only legitimate transactions are added to the blockchain and that the network is kept safe. BTC has a limited supply of 21 million coins, ensuring that its inflation rate is predictable and controllable. In recent years, it has seen widespread attention, not only in terms of adoption as a payment method but also in trading. Therefore, accurately predicting its price provides a unique advantage to the trader.

Predicting the price of BTC is an interesting and important task for a variety of reasons. One reason is that it can help investors and traders make educated decisions about whether to buy or sell BTC. This task is also vital for companies and individuals that accept it as a payment method, as it allows them to plan for anticipated fluctuations in the value of the currency [2,3]. Understanding the variables driving the price of BTC can also help

politicians and regulators decide how to handle the cryptocurrency market [4–6]. In general, projecting the price of BTC can provide useful information on the current situation and future possibilities of the cryptocurrency sector [7,8]. Furthermore, it has been shown in several independent studies that adding other attributes in addition to the price to a given forecasting model can aid its forecast accuracy [9–11]. A promising approach is to use sentiment scores derived from the content of news articles, blogs, or microblogs that correspond to the same period as the price values [12–14]. For example, Vyas et al. [15] and Zhou et al. [16] show the potential of BERT-based sentiment analysis models for contemporary social problems related to the Russia–Ukraine war or for identifying rescue request tweets, respectively. In accordance with these studies, we use sentiment analysis techniques to predict the price of Bitcoin. Many individuals have invested considerable amounts of their personal wealth in assets such as stocks, gold, or, recently, cryptocurrencies. The latter, including Bitcoin, tend to be particularly volatile. Therefore, having a sound estimate of the market sentiment and creating forecasts based on such sentiments can help these individuals avoid detrimental consequences. For instance, the cryptocurrency market, including Bitcoin, crashed by 25% in less than 48 h following the first signs of the collapse of the popular FTX exchange on 9 November 2022, leading to unexpectedly high wealth depletion [17,18]. In turn, the cryptocurrency market sentiment plummeted [19]. The availability of a sentiment-enriched forecasting model could have alleviated this immense financial damage to many individuals.

Sentiment analysis (SA) techniques are particularly useful for predicting the price of assets since they provide insights into the public's perspective and sentiment about the asset. This data can give useful insights into the general market mood, which has been proven to be a vital determinant in financial decision making [20–22]. Research has shown that people make judgments based on emotions rather than logic [23]. Emotional decision making can be captured through SA in social media data, which can then be used to forecast prices. Another promising area of research is emotion detection (ED), which differentiates itself from SA by classifying a specific set of emotions [24]. Furthermore, psychological research has indicated that investor emotion is an important factor in financial decision making. For example, Barberis et al. [25] demonstrated that the investor attitude may influence asset prices and be used to forecast future returns. Similar findings have been made for cryptocurrencies, including BTC [20]. As a result, SA can be a useful method for predicting BTC prices, since it provides insight into the public's view and mood toward the cryptocurrency.

Traditionally, such techniques have relied on a complex set of rules. One popular method that has been designed for microblogs is VADER [26]. Although it has shown success in deriving accurate sentiment scores [10,26–28], it also has major shortcomings. Given its rule-based approach, it cannot capture complex relations between individual words. This, however, is where neural methods excel. A popular model that has been used successfully for all types of natural language processing (NLP) tasks is BERT [29]. Recently, many fine-tuned variants of BERT have been introduced, including versions specifically tuned for microblogs and SA, as well as ED [30] tasks. Since they have also been shown to excel in such tasks, our aim in this study is to make use of the capacity of such language models when integrating them into different forecasting models.

In this article, our aim is to forecast the price of Bitcoin with the help of sentiment scores derived from SA techniques. Against this background, we hypothesize that adding collaborative sentiment information inferred from tweets to time series forecasting improves the accuracy of predictions. The main contributions of this work are the following:

- We propose a new hybrid method that integrates time series forecasting and sentiment analysis based on a fine-tuned BERT model, featuring a novel weighting scheme to aggregate multiple sentiment scores from a given period into a single sentiment score.
- We thoroughly investigate our approach, which spans, compared to previous research, both longer and more diverse price ranges and market scenarios, making the task more realistic but also much harder.

- Using this setup, we show that our approach outperforms previous ones. In particular, we empirically show that both our BERT model fine-tuned for sentiment analysis and our novel weighting scheme improve forecasts in terms of predictive accuracy (MAE and RMSE) compared to other setups. Moreover, we show that simpler models, particularly linear regression models, tend to perform best, while more complex models have issues with overfitting.

In the remainder of the article, we first give an overview of comparable approaches (Section 2). Then, we specify our methodology in detail (Section 3), first in terms of forecasting algorithms (Section 3.1) and then in terms of sentiment analysis methods (Section 3.2) and also in terms of our data setup (Section 3.3). Finally, in Section 4, we share our empirical results and discuss them both qualitatively and quantitatively. An overview of our approach is also given in Figure 1.

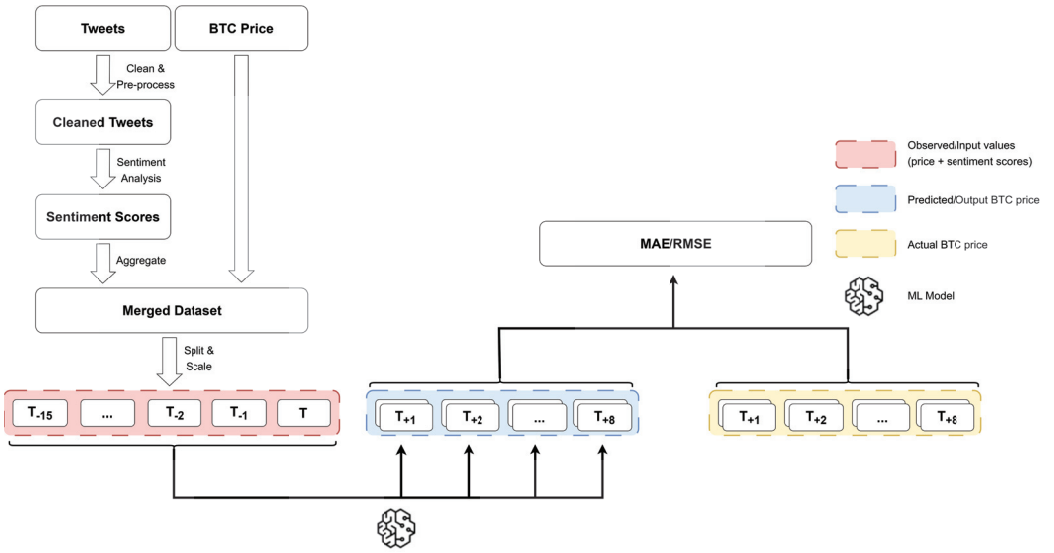


Figure 1. Overview of our BTC price forecasting pipeline. After having collected and pre-processed the data, we feed 16 past BTC price values and sentiment scores into our forecasting models, which generate forecasts for the future eight time steps. We compare our models with the actual BTC price in terms of MAE and RMSE error metrics.

2. Related Work

Our work connects to two strands of research in the domain of Bitcoin price prediction: time series forecasting (Section 2.1) and sentiment analysis (Section 2.2). State-of-the-art methods combining both areas are reviewed in Section 2.3. We provide a summary table of research work related to Bitcoin price prediction with the help of sentiment scores in Table 1. We also sketch our novel contributions in this work (Section 2.4).

2.1. Time Series Forecasting for Cryptocurrencies

To obtain forecasts, there exist numerous models, ranging from very simple to highly complex [9,31]. One strand of research employs deep learning models, such as long short-term memory (LSTM) networks, to anticipate the price of BTC. These models have been successful in capturing temporal relationships in data and can be trained on vast volumes of data [31–33]. LSTM networks have also been successfully applied to predict the price of BTC [2,34–36].

Jaquart et al. [34] trained a diverse set of forecasting models to predict binary relative daily market movements of 100 cryptocurrencies, including Bitcoin, and showed that an ensemble of LSTM and gated recurrent unit (GRU) networks, both of which belong to the category of recurrent neural networks (RNNs), performs best. Uras et al. [35] utilized multi-layer perceptrons (MLPs), LSTM networks, and linear regression models to predict the price of BTC, along with the prices of other stocks, and showed that LSTM

networks performed best. Patel et al. [2] and Hamayel and Owda [36] used LSTM and GRU networks to forecast the price of BTC and also that of other cryptocurrencies. Their experimental results indicate that models that perform well on BTC perform comparatively well on other cryptocurrencies, such as Ethereum, Litecoin, or Monero. Contemporary to our study, Dimitriadou and Gregoriou [37] collected 24 relevant variables, such as exchange rates or the value of other cryptocurrencies, to predict whether BTC would rise or fall. In their results, the simple logistic regression model performed the best. In another research work, three models, namely, an MLP, a support vector machine (SVM), and an LSTM network, were trained to predict the price of BTC using its historical price and several additional BTC trading indicators as input features [9]. The authors found that the SVM with comparably few trainable parameters outperforms all others, including the more complex ones. However, they failed to employ a proper, simple baseline model that does not consider any additional features other than the historical price of BTC. Moreover, Liu et al. [31] drew two conclusions that provide valuable insights relevant to our research. First, they found that, in various time series forecasting tasks, more complex models often suffer from overfitting, which they attributed to the limited availability of training data. Second, they contended that transformer networks exhibit promising results in certain time series forecasting tasks but are computationally expensive and unable to effectively capture long-term dependencies from the distant past.

2.2. Sentiment Analysis

Another strand of research aims to capture sentiment scores derived from SA techniques. In general, SA can be categorized into machine learning (ML), lexicon-based methods, and rule-based techniques [38,39]. For text found on social media platforms, one prominent method is VADER, which we describe in more detail in Section 3.2.1. VADER has been employed for a variety of tasks; for example, for tasks specific to the COVID-19 pandemic [40,41], for app recommendations [42], or for classifying complex datasets [43]. Another SA method commonly used for social media data is TextBlob [44]. In contrast to VADER SA, which derives its sentiment scores based on rules, TextBlob combines lexicon-based approaches and ML techniques [45,46]. Notably, TextBlob has also been used for emotion detection by Aslam et al. [47]. More recently, SA methods based solely on ML have gained popularity [48], for example, methods based on convolutional neural networks [49], recurrent networks [47], attention networks, [50] or transformer networks [51]. We provide more details on the latter (in particular, BERT) in Section 3.2.2.

2.3. Combining Time Series Forecasting with Sentiment Scores

In addition to the historical price of financial assets, sentiment scores have been added as an additional input feature to the model (most commonly, LSTM networks), in a number of studies, with encouraging results [52–54]. Furthermore, several studies have shown the ability of sentiment scores derived from the VADER SA technique to enhance the prediction of the price of BTC [10,54–56]. For instance, Abraham et al. [57] and Mittal et al. [58] combined sentiment scores with tweet volumes and Google Trends, respectively, to create forecasts. Moreover, Wołk [14] used ten different forecasting models to predict the price of several cryptocurrencies, including Bitcoin, and showed that psychological and behavioral aspects have a strong impact on prices. In a recent study, Ye et al. [54] created a stacking ensemble consisting of LSTM and GRU networks to predict the price of BTC based on historical price data, various technical indicators, and VADER-derived sentiment scores. Their results indicate that the combination of multiple model outputs through stacking can improve forecast accuracy and that both technical indicators and sentiment scores improve forecast accuracy. Likewise, Serafini et al. [13] employed two distinct models in their research, namely, an ARIMAX model and an LSTM model, to forecast BTC prices. The authors conducted experiments using various combinations of sentiment and financial input features. In particular, their findings revealed that the inclusion of additional input features, such as the BTC trading volume or the tweet volume, did not necessarily improve

the forecasting capabilities. This observation, which may seem counterintuitive for this specific task, served as a key motivation for our study. Our aim is to investigate the analysis of optimal combinations of characteristics and explore possible adjustments to improve the performance of forecasting models.

Furthermore, they found that although the ARIMAX model they employed is simpler it outperformed the more complex LSTM model. In another study, Edgari et al. [59] used XGBoost, an implementation of a gradient-boosted decision tree, to estimate the price of BTC during the COVID-19 pandemic. They used a dataset comprising tweets that date from February 2021 to December 2021 and aggregated tweets into 1-min time windows. To obtain sentiment scores, they applied VADER SA and used the average of the compound score as the sentiment score for a given 1-min time interval. In addition to sentiment scores, they added the average of user-related features over all users who created a tweet in a given 1-min interval as input to the model, such as the number of followers, favorites, and whether the user is verified or not. Finally, they predicted whether the price associated with the next time step rises or falls, thus creating a binary classification task. Using the aforementioned XGBoost model to make predictions, the authors concluded that sentiment scores are indeed helpful for forecasting BTC prices. A summary table of the research works related to the prediction of BTC prices with the help of sentiment analysis techniques is shown in Table 1.

Table 1. Summary of related work on BTC price prediction leveraging sentiment scores. *Year* corresponds to the publication year. *Pred. target* refers to the forecasting target of the BTC price: *binary* means that the objective is to predict whether BTC will increase or decrease, and *value* means that the objective is a direct prediction of the BTC price. *Granularity* corresponds to the degree to which the data are aggregated for prediction. *Model* corresponds to the set of forecasting models used. *Additional features* lists the features that are added in addition to the historical BTC price.

Ref.	Year	Pred. Target	Granularity	Model	Additional Features
[55]	2020	binary	1 min	random forest	VADER SA
[54]	2022	value	30 min	stacking ensemble of LSTM, GRU	VADER SA, trading features
[56]	2022	value	12 h	vector autoregression	VADER SA
[57]	2018	value	1 day	linear regression	VADER SA, tweet volume, Google Trends
[58]	2019	value	1 day	LSTM, vanilla RNN, linear regression, polynomial regression	VADER SA, tweet volume, Google Trends
[14]	2019	value	1 day	10 different models	VADER SA, Google Trends
[13]	2020	value	1 day	LSTM, ARIMAX	VADER SA, tweet volume
[59]	2022	binary	1 min	XGBoost	VADER SA, tweet volume, user-related features

2.4. Research Gaps and Novel Contributions

Reviewing related work, we identified research gaps concerning (1) the period spanned by the investigated datasets and the corresponding BTC price ranges, (2) the use of more up-to-date and sophisticated SA and ED techniques, and (3) the use of appropriate baselines. Furthermore, we have not encountered studies that predict not only one but multiple time steps into the future. Addressing these gaps, what sets the article at hand apart from previous research is that we use a dataset spanning a longer period and more diverse price ranges. Moreover, we employ not only VADER SA but also BERTweet-based SA and ED. Additionally, we utilize a distinct set of forecasting models, including appropriate baselines. The dataset we use comprises tweets from February 2021 up to November 2022, which makes it more recent and covers a longer period as well as more diverse price ranges compared to the one used by Edgari et al. [59] and other previous studies. Furthermore, we conduct our investigation over a 30-min period rather than a 1-min period, as employed by

Edgari et al. [59]. This enables us to gain a more complete understanding of the market sentiment and its influence on the price of BTC over time. Moreover, instead of the more conventional way of predicting the price of a given asset in a time step p_{T+1} using the prices of $p_T, p_{T-1}, ..., p_{T-N}$, where N may vary, we predict M time steps, with $M > 1$, thus predicting the price of BTC for multiple time steps into the future. Although more challenging, this approach allows for a longer forecast period, ranging from p_{T+1} to p_{T+M}. Consequently, it integrates the performance assessment of short- and mid-term forecasts using just a single unified model. Predicting multiple time steps into the future enhances the forecasting horizon, offering flexibility and adaptability in decision making to align with different potential future scenarios. It also allows for the assessment of the model's performance across different time frames, contributing to a more holistic understanding of market dynamics and providing insights into the model's stability over time. Moreover, to the best of our knowledge, we are the first to use BERT-based sentiment analysis techniques to enrich the historical price of BTC with sentiment scores, along with emotion detection scores, combining these unique forecasting characteristics.

3. Methodology

In the following, we provide a brief overview of the overall task as well as our approach. The exact methodology of our hybrid BTC price forecasting algorithm will be explained in more detail in the upcoming sections.

As visually described in Figure 2, we first gather data on both the historical price of BTC and the time-aligned tweets, from which sentiments are extracted. The textual tweet dataset is then pre-processed using standard NLP techniques, cleaned, and subjected to three distinct types of SA methods. Next, we generate sentiment scores for each tweet. For that, we use two distinct approaches, namely, VADER, which relies on rules, and different variations of a BERT model, which is an ML model that can capture the relations between words more richly compared to traditional rule-based methods. We group the output into bins of 30-min intervals based on the corresponding tweets' time stamps and then use two distinct methods to weigh the derived sentiment values. The last step of data pre-processing is to prepare the data so that it can be used as input to various forecasting methods.

Finally, as can be visually observed in Figure 1, we feed the prepared data into our ML models. We use 16 30-min intervals, that is, 8 h, as input and predict 8 30-min intervals, that is, 4 h, for all our forecasts. To predict the future price of BTC, we employ four different types of models: linear regression (LR), LSTM networks [60], temporal convolutional networks (TCNs) [61], and the D-Linear method [62], each of which uses a unique set of covariates.

In the next sections, we outline the forecasting algorithms used (Section 3.1), specify our method to compute sentiment scores (Section 3.2), and outline the data acquisition procedure, as well as the processing steps (Section 3.3).

3.1. Forecasting BTC Price

We conduct most of our forecasting experiments with the *Darts* [63] framework, which is based on PyTorch [64]. We feed 16 30-min intervals, i.e., 8 h, as input to our models and predict 8 30-min intervals, i.e., 4 h, for all our forecasts. These values are chosen to reflect the characteristics of intraday trading. We evaluate the final performance on the test set, which has not been considered at all until the final evaluation. We use the validation set to select the best hyperparameter setup for each model, including early stopping. More information on how we split the data is given in Section 3.3.

Forecasting Algorithms. In our hybrid BTC price forecasting algorithm, we investigate four types of ML-based algorithms: LSTM networks, TCNs, the D-Linear method, and linear regression, along with some simple baselines. We decide to use LSTM networks and TCNs since both have shown remarkable success in a variety of forecasting tasks [14,34,53,65,66], making them a very common choice for models with high complexity. Moreover, we decide to use the D-Linear method and linear regression

since both have shown potential in making accurate forecasts while being comparatively simple [35,62,67,68]. For forecasting, all our baselines rely solely on the price and do not require any training. The other models that are employed, except for the LR model, are all trained with the following commonalities: We train each of them for 50 epochs using the mean squared error (MSE) loss function and the Adam optimizer, but we use early stopping if the loss on the validation set does not decrease for five consecutive epochs. Additionally, if the training loss does not decrease for three successive epochs, we cut the training rate by half. More details on hyperparameter selection are provided in the following subsections that detail the algorithms used.

3.1.1. Baselines

Exponential Smoothing. Exponential Smoothing (ES) is a time series forecasting method that uses a weighted average of past observations to predict future observations [69]. The weights decrease exponentially as observations come from further in the past; the smallest weights are associated with the oldest observations. It is a generalization of the simple moving average, where all weights are equal. We use Holt–Winters' ES [70] with additive seasonality as well as trends and consider two periods in each seasonal cycle.

Fast Fourier Transform. The Fast Fourier Transform (FFT) decomposes the time series into its frequency components [71,72], which are then used to predict future values of the time series. It is particularly suitable for highly seasonal data. We use the standard *Darts* hyperparameters.

Naive Mean. The Naive Mean model simply predicts the mean value of the input series. Hence, in our case, it predicts the mean value of the last 16 BTC price values.

Naive Drift. The Naive Drift Model extends the line between the first and last point of the training series into the future [63].

3.1.2. Long Short-Term Memory Network

The long short-term memory (LSTM) network, as introduced by Hochreiter and Schmidhuber [60], is a type of recurrent neural network used for sequence prediction problems. The network has memory, which allows it to learn from sequences of input data. The LSTM model consists of memory blocks that are linked together. Each block has a memory cell that is connected to gates that control the flow of information into and out of the memory cell. The LSTM model is trained by unrolling the network and presenting the input sequence one element at a time to the network. LSTM networks have been used successfully for a variety of tasks, including time series forecasting [32,33]. To predict multiple time steps into the future, the LSTM network learns a single-step forecast and iteratively applies it to obtain multi-step forecasts (in our case, 8). Hence, it can be characterized as autoregressive.

We train LSTM networks with a batch size of 2048, a learning rate of 0.0003, and a dropout rate of 0.1. Moreover, we use LSTM networks with two layers with a hidden dimension of size 512 and append two fully connected layers with sizes $[512, 128]$ to the hidden layers.

3.1.3. Temporal Convolutional Networks

Temporal convolutional networks (TCNs) are a class of neural networks that are designed to operate over sequences of data, such as time series [61,73]. TCNs consist of a stack of convolutional layers, pooling layers, and fully connected layers. The convolutional layers are comprised of filters (kernels) that are applied to the input sequence; pooling layers are used to down-sample the output of the convolutional layers; and the fully connected layers are used to interpret the features extracted by the convolutional and pooling layers.

We train TCNs with a batch size of 512, a learning rate of 0.001, and a dropout rate of 0.2. We use TCNs with 16 kernels of 8, 12 layers, and 2 as a dilation base. Moreover, we apply weight normalization to stabilize the training process.

3.1.4. D-Linear

The D-Linear model is a novel time series forecasting method based on a single-layer neural network [62]. First, it decomposes the input data into a trend by applying a moving kernel average and a remainder (seasonal) component. Second, two single-layer linear layers are applied to each component. Ultimately, the two features are summed up to obtain the final predictions. Therefore, it has comparatively few parameters.

We train the model with a batch size of 2048, a learning rate of 0.01, and no dropout. Moreover, we use a kernel size of 25 for the moving kernel average computation.

3.1.5. Linear Regression

Linear regression is a statistical method for modeling the relationship between a dependent variable y and one or more explanatory variables (or independent variables) denoted by X [74]. Since we rely on more than one explanatory variable, we conduct multiple LR models. We perform least squares regression to find the best-fitting line for the data, which is the line that minimizes the MSE between the predicted and target values. To generate predictions for multiple time steps, we train a separate model for each future time step. This results in eight independent models for our approach, each corresponding to a single future time step.

3.2. Sentiment Analysis

The purpose of sentiment analysis, a branch of natural language processing, is to discover the general view or attitude of a speaker or writer toward a certain topic or issue automatically [38,39]. Machine learning, lexicon-based methods, and rule-based approaches are some of the techniques that may be utilized for SA [38,39].

In our study, we rely on two distinct SA approaches: The first is VADER, which relies on a set of rules and has been used successfully in many studies [10,55,56,59]; but it cannot capture the relations between words. It is specifically designed for social media data. The second relies on BERT, which has shown great promise in various NLP tasks, including SA and ED. Since BERT is a general-purpose model, we use fine-tuned variants of BERT that suit our hybrid BTC price forecasting algorithm.

3.2.1. VADER Sentiment Analysis

VADER is an SA method that has been specifically designed for social media data [26]. The rule-based model has become one of the most popular since its introduction in 2014 [10,11,59,75]. To define these rules, the authors introduce a combination of qualitative and quantitative methods to generate a "gold-standard" lexicon that is specially tuned for social media text, all while following an "explicit human-centric approach". Hence, each word in the English language is associated with a sentiment *polarity* (positive/negative) and sentiment *intensity*. These scores are aggregated to obtain scores for a single passage or document. Finally, upon qualitative analysis, the authors decide to use further heuristics incorporating word-order sensitive relationships between terms, namely, punctuation, capitalization, degree modifiers, "but" as a sentiment polarity shifter, and catching negation flips by considering tri-grams. These heuristics modify the score of a given passage or document. For each piece of text, VADER outputs four kinds of sentiment scores: *positive, negative, and neutral*, corresponding to sentiment *polarity* with a given *intensity*, and a *compound* score. For the SA methods, we conduct experiments using *all* scores but also only using the *compound* score as input to the model in addition to the historical price data.

To obtain sentiment scores using VADER, we use the open-source Python implementation provided by the authors of the original paper [76].

3.2.2. BERT-Based Sentiment Analysis

BERT is a pre-trained transformer model that has been successfully applied to a wide variety of tasks and domains [29]. The attention mechanism employed in these models

allows them to capture the context, specifically the surrounding words in the sentence. We rely on two different, fine-tuned versions of BERT, each of which outputs a distinct set of sentiment features. Both rely on BERTweet [30], which is a version of BERT that has been further trained on English tweets.

Despite being much more computationally expensive than VADER, we employ BERTweet-based SA and ED models, since both models have shown state-of-the-art performance on their respective tasks [30]. Moreover, they have not yet been used to predict the price of BTC to the best of our knowledge.

To obtain sentiment scores using BERTweet-based SA and ED models, we use the respective models for English, as provided by the open-source Python library *pysentimiento* [51].

BERTweet-based sentiment analysis. The first specialized model is a BERTweet-based SA model for English tweets. It starts from the weights of BERTweet and further fine-tunes them on the SemEval-2016 dataset [77], which is an SA dataset. For each piece of text, it produces three types of sentiment scores: *positive*, *negative*, and *neutral*. In addition, we define a compound score, which simply consists of subtracting the negative score from the positive score.

BERTweet-based emotion detection. The second specialized model is a BERTweet-based ED model for English tweets. Again, it starts from the weights of BERTweet, but this time it is further fine-tuned on the EmoEvent dataset [78], which is an ED dataset. For each piece of text, it produces seven scores, each corresponding to a distinct emotion: joy, sadness, anger, surprise, disgust, fear, and, additionally, a score for *others*. We have decided to also include ED since emotions enable a more nuanced distinction of sentiment towards BTC compared to only using *positive*, *negative*, and *neutral* scores, which also improves interpretability. In this setup, we use all seven emotion scores as input to the model, in addition to the historical price data.

3.3. Data Acquisition and Processing

In the following, we outline our data acquisition and processing approach. First, we describe how we collect our data (Section 3.3.1) and how we pre-process the tweets (Section 3.3.2). Then, we explain how we aggregate multiple sentiment scores of tweets from a given time interval into a single sentiment score for the given time interval (Section 3.3.3). Finally, we show how we merge our two data sources and how we split the data into train, validation, and test sets (Section 3.3.4). Our data and pre-processing pipeline is shown in Figure 2.



Figure 2. Overview of our data acquisition and pre-processing pipeline. After having collected both BTC prices and relevant tweets, we pre-process the text, resulting in cleaned tweets. Next, we derive sentiment scores using VADER SA, BERTweet SA, and BERTweet ED. Based on these sentiment scores, we aggregate multiple tweets using two different weighting schemes. Finally, we merge the two datasets and split the data into train, validation, and test sets. Previous methods are marked in *blue*, whereas our introduced methods are in *red*.

3.3.1. Data Collection

For our evaluation experiments, we require some historical BTC price data as well as tweets. We leverage the *Binance* API [79], which offers price data at 30-min intervals, to obtain the information for the BTC price. We use the BTC-USDT pair, as USDT (Tether) is the stablecoin with one of the highest market caps tied to the US Dollar, which in turn is the most significant currency for worldwide trading. Hence, this provides a realistic scenario from the point of view of a trader.

Secondly, to obtain sentiment scores that are calculated based on tweets from *Twitter*, we use a dataset from *Kaggle* [80], which consists of tweets with the hashtag #BTC or #Bitcoin. Furthermore, the dataset consists of tweets spanning from 10 February 2021 to 20 November 2022. This period saw particularly volatile BTC prices. Later, forecasting these prices from these diverse ranges proved to be a significant challenge. The dataset consists of 4.5 million relevant tweets in total, with 11 attributes each. For our use case, only the attributes "text" (i.e., the content of the tweet), "date" (the exact time of creation), and "user_followers" (i.e., how many followers the author of the tweet had at the time of posting) are of relevance. However, the dataset comes with one limitation: there are numerous temporal gaps—periods when there were no tweets at all. They frequently lasted for several weeks and, in fact, more than half of the days lack any tweet data. Therefore, we had to discard these time periods, which also significantly reduced the amount of data on which we could reliably train our models. However, we still opt to use this dataset, as it is quite large. Moreover, one major goal of this study was to see how well models generalize over unseen price ranges and market conditions; for that, we needed a dataset that spans periods corresponding to varying market conditions. This also clearly differentiates our setup from those used in previous studies and makes it more useful for real-life use cases.

3.3.2. Text Pre-Processing

The dataset is first cleaned using common practices as employed in the literature as part of our text pre-processing process [12,56,59]. We eliminate non-English tweets with *langdetect* [81] and tweets with fewer than three characters. We also replace over two (or more) subsequent dots with a space, remove multiple spaces, and remove words with only one character with the Python-based open-source library *Pandas*. After having cleaned the dataset, 458,472 tweets remain for further pre-processing. Subsequently, we prepare the text in a format with which our sentiment analysis methods work well. This process is different for our different approaches. To prepare the data for VADER, we lowercase the tweets; remove @, $, and # signs; and remove stopwords. Although both methods have been used in the related literature [12,56], we choose not to lemmatize or stem the text because in our dataset there are many domain-related terms, abbreviations, and jargon that may not necessarily benefit from common lemmatization or stemming techniques. To prepare the data for BERT-based SA, we follow the process described by Pérez et al. [51]: # signs, @ signs, and URLs are replaced with special tokens (and not removed, as performed for VADER), and emojis are converted to their text representations.

3.3.3. Data Aggregation

Having sentiment scores for each tweet, the next step is to aggregate the sentiment of all tweets from a given 30-min interval into a single sentiment score corresponding to the overall sentiment in a given time period (in our case, 30-min). There exist different ways to weigh the scores, such as taking the mean, median, or sum of all the sentiment scores. We conduct experiments with two different weighting schemes. For each of them, we independently weigh the different scores of each SA method.

Mean weighting. The first is to simply take the mean sentiment score of all tweets in each 30-min interval. In other words, we compute

$$\sum_{i=1}^{n} \frac{s(x_i)}{n},$$

where $s(x_i)$ is the sentiment score of a tweet x_i in the current 30-min interval, and n is the number of tweets in this interval.

Weighting by number of followers. The second one, which we will refer to as *user influence* weighting, relies on the assumption that Twitter users with more followers are

more influential and therefore presumably more relevant. Based on this, we arrive at the following novel weighting scheme:

$$\frac{\sum_{i=1}^{n} s(x_i)(\ln(w_i+1)+1)}{\sum_{i=1}^{n} \ln(w_i+1)+1},$$

where $s(x_i)$ is the sentiment score of a tweet x_i in the current 30-min interval, and w_i is the number of followers the author of the tweet had at the time of creation. We use the logarithm to dampen the effect of users with a very large number of followers, inspired by the retrieval method BM25 [82].

3.3.4. Data Source Merging and Splitting

Finally, we merge the two data sources, namely, historical BTC prices and sentiment scores from *Twitter*, into a single data source, which can be neatly achieved by matching their respective time steps. Next, we split the data into train, validation, and test splits based on time to counteract the "peek-a-boo" anti-pattern [83], which would be violated if we used a random or stratified split based on the target value, an otherwise common practice in ML. We used 80% of the available data for training and 10% as validation and test sets, respectively. Using this splitting strategy, the training data spans from 11 February 2021 up to 6 June 2022; the validation data from 14 June 2022 up to 6 September 2022; and the test data from 11 September 2022 up to 20 November 2022. The gaps in the respective sets are due to the gaps in the dataset because we want to avoid splitting during a period with a gap. Note that the size of the splits does not depend on the date but rather on the number of 30-min intervals available during a given time interval. Finally, we standardize the data based on the values in the training set.

We also experimented with additional feature engineering, including exponential moving averages, temporal features such as the day of the week or holidays, and the total number of tweets in each time interval (i.e., tweet volume). However, none of these significantly improved performance metrics in our experiments. We assume that this may be due to the highly volatile nature of BTC.

4. Evaluation and Discussion

In this section, we present the results of our main experiments, which aim to evaluate the performance of the investigated BTC price forecasting algorithms. First, we provide a comprehensive overview of the performance of all algorithms, including baselines, when only considering the price as input. Next, we compare the performance of all of the algorithms that we used with varying input features, using the mean weighting scheme. This gives us a better understanding of how each algorithm performs with different sentiment scores. Afterward, we compare the performance of the mean weighting scheme with our user influence weighting scheme. We demonstrate how our user-influenced weighting scheme affects the overall performance of the algorithms and identify if it outperforms the mean weighting scheme. Overall, these experiments provide us with a better understanding of how each algorithm performs under various conditions and help us determine the most efficient algorithm and weighting scheme for our study. By analyzing the results of these experiments, we are able to identify the most suitable algorithm and combination of features to predict the price of BTC.

For reference, the minimum price of BTC during the entire period was USD 15,704, while the maximum value was USD 68,633, with the latter occurring during the training set period. However, the minimum price of the training set was USD 27,114, whereas the minimum price occurring during the test set period was USD 15,704. The maximum price that occurred during the period of the test set was USD 22,713. This divergence highlights the difference between the training and test set price ranges. Therefore, in order to perform well on the test set, our models need to have good generalization capabilities.

We evaluate all our models on the test set using the RMSE (root mean squared error) and MAE (mean absolute error) error metrics. Since we compute multiple forecasts, we

consider every output prediction, compare it with the ground truth, compute the respective error metric over our eight forecasting time steps, and then take the average, as can be seen in Figure 1. Then, we move one time step forward and again compute the same metric over the next eight forecasting time steps. The final results, as reported in Tables 2–4, are obtained by computing the arithmetic mean of the averages of the computed error metrics.

To investigate whether our best models are significantly better than the other models, we make use of statistical testing. To accomplish this, we first test whether our dataset is stationary or normally distributed, since some commonly used tests, such as ANOVA, require normally distributed samples. To check whether the data are normally distributed, we use the Shapiro–Wilk test. The test reveals that the dataset does not follow a normal distribution. We also employ a quantile–quantile (qq) plot to visually check the data to verify our findings. Moreover, to check stationarity, an important aspect, we conduct a Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test. On the basis of this test, we conclude that Bitcoin price data are non-stationary, which might be associated with their highly volatile nature in general. Moreover, we use the *statsmodels* Python library to split the price data into seasonal, trend, and residual components. Based on these components, we find that the time series does not display a significant seasonal pattern. Moreover, we notice a declining tendency in the validation set but not in the training set. Finally, we validate the time series by plotting the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the BTC price. In summary, the results of the ACF plot (e.g., 0.5 at lag 20) indicate a moderate correlation between the current value and its past values up to lag 20. However, the PACF plot indicates a very strong correlation of almost 1 of the first 2 lags with the current value, while lags from the third and onward show a correlation closer to 0, suggesting that the current value becomes less dependent on its past values beyond the immediate two lags. By conducting these tests and analyzing the results, we gained valuable insights into the properties of the Bitcoin datasets, which will inform our subsequent modeling and analysis. Furthermore, it helps in choosing the appropriate statistical test. Based on our findings, using the results on the test set of all models, we conduct a permutation test, as it is more robust w.r.t. different sample distributions and, hence, more robust than the commonly used t-test. Using this test, we use a high permutation number of 1,000,000 to obtain a higher precision of the p-value, resulting in a standard deviation of only 1.2×10^{-4}. In the following, we discuss a few observations, both in terms of models and features.

Table 2. Comparison of performance of forecasting algorithms on the test set using *only* the BTC price as input, with data spanning from 5 February 2021 up to 20 November 2022. The best score for a given feature combination is marked in **bold**, and the second best is underlined. Columns on the right indicate the different forecasting algorithms, as specified in Section 3.1. All of them use the BTC price as the only input attribute. We conduct a permutation test to determine statistical significance between our best models and the other models. † indicates that all of the LR, D-Linear, ES, and Drift models perform significantly better ($p < 0.05$) than all other models with worse performance.

Metric	LSTM	TCN	D-Linear	LR	ES	FFT	Drift	Mean
MAE	11.69	13.71	3.40 †	**2.72** †	4.15 †	9.46	3.43 †	7.13
RMSE	12.22	14.19	4.22 †	**3.33** †	5.01 †	10.04	3.97 †	7.51

Table 3. Comparison of performance of algorithms and SA techniques on the test set using *mean weighting*, with data spanning from 5 February 2021 up to 20 November 2022. The best score for a given feature combination is marked in **bold**, and the second best is underlined. Columns on the right indicate the different forecasting algorithms, as specified in Section 3.1. Rows indicate the covariates used as inputs to the different forecasting algorithms, depending on the SA method employed. *All scores* means that all the sentiment scores are used as covariates, while *compound score* corresponds to only using the compound score that is calculated based on the other scores as a covariate. All setups use the past price of BTC, as described in Section 3, as input. We conduct a permutation test to determine statistical significance between our best models and the other models. † indicates that both the LR and D-Linear models perform significantly better than the other models, while ‡ indicates that our LR model is better than all other models with worse performance, including D-Linear. Both † and ‡ correspond to $p < 0.05$.

SA Method	Scores	Metric	LSTM	TCN	D-Linear	LR
VADER SA	All	MAE	90.50	19.22	4.49 †	**3.23** †
		RMSE	90.67	19.69	5.48 †	**3.96** †
	Compound	MAE	25.81	12.84	3.46 †	**2.71** †
		RMSE	26.17	13.37	4.23 †	**3.31** †
BERTweet SA	All	MAE	53.32	22.72	4.75 †	**3.14** †
		RMSE	53.09	22.18	5.93 †	**3.85** †
	Compound	MAE	35.04	11.56	3.75 †	**2.67** †
		RMSE	35.41	12.10	4.75 †	**3.28** †
BERTweet ED	All	MAE	22.39	15.49	6.35	**2.87** ‡
		RMSE	22.94	16.07	8.02	**3.52** ‡

Table 4. Comparison of performance of algorithms and different weighting schemes on the test set using *all sentiment scores*, with data spanning from 5 February 2021 up to 20 November 2022. The best score for a given feature combination is marked in **bold**, and the second best is underlined. Columns on the right indicate the different forecasting algorithms, as specified in Section 3.1. Rows indicate the covariates used as inputs to the different forecasting algorithms, depending on the SA method employed. *Mean weighting* and *user influence weighting* refer to the different weighting schemes introduced in Section 3.3.3. We conduct a permutation test to determine statistical significance between our best models and the other models. † indicates that both the LR and D-Linear models perform significantly better than the other models, while ‡ indicates that our LR model is better than all other models with worse performance, including D-Linear. Both † and ‡ correspond to $p < 0.05$.

SA Method	Weighting	Metric	LSTM	TCN	D-Linear	LR
Vader SA	Mean	MAE	90.50	19.22	4.49 †	**3.23** †
		RMSE	90.67	19.69	5.48 †	**3.96** †
	User influence	MAE	132.54	25.55	4.37 †	**3.35** †
		RMSE	132.66	25.98	5.36 †	**4.11** †
BERTweet SA	Mean	MAE	53.32	22.72	4.75 †	**3.14** †
		RMSE	53.09	22.18	5.93 †	**3.85** †
	User influence	MAE	53.09	10.13	4.53 †	**3.01** †
		RMSE	53.33	10.78	5.68 †	**3.69** †
BERTweet ED	Mean	MAE	22.39	15.49	6.35	**2.87** ‡
		RMSE	22.94	16.07	8.02	**3.52** ‡
	User influence	MAE	36.13	14.43	5.97	**2.84** ‡
		RMSE	36.48	14.95	7.36	**3.49** ‡

4.1. Model Comparison

First, we compare our different models with each other: LR, D-Linear, TCN, and LSTM. The results are shown in Table 3.

Models with a lot of parameters (LSTM and TCN) are prone to overfitting, which is a typical problem that occurs when a model is overly sophisticated and can recall the training data but fails to generalize to new data, i.e., test data. Indeed, they perform very well on the training set but perform poorly on the test set. Specifically, the variance in their predictions on the test set is very high, and their predictions are generally too high. Increasing regularization, such as dropout or weight decay, did not ameliorate their weak performance. We hypothesize that this may be due to not using enough training data and that the gaps in the training set in particular hurt their generalization performance significantly. Second, the D-Linear model performs decently but not as well as LR. Compared to the baselines, it is only better in terms of the MAE. As an ultimate observation, we see that LR is by far our best model. It does not have issues with overfitting and is also the fastest to train. We hypothesize that this may be due to the unconventional method we have chosen, i.e., fitting a single LR model for each output prediction.

Figure 3 shows the performance of our models on the last segment of the test set using *mean weighting* and *all sentiment scores*. The results indicate that the more complex models in particular have difficulties with previously unseen price ranges and market conditions observed in the test set periods. The predictions of the complex LSTM and TCN models are rather unstable and always overestimate the true price, indicating that they are unable to generalize. In contrast, the simpler D-Linear model predicts the correct price ranges more accurately but is still unstable. Finally, the LR model predicts the general price ranges well while also making the best forecasts, showing its superior generalization capabilities.

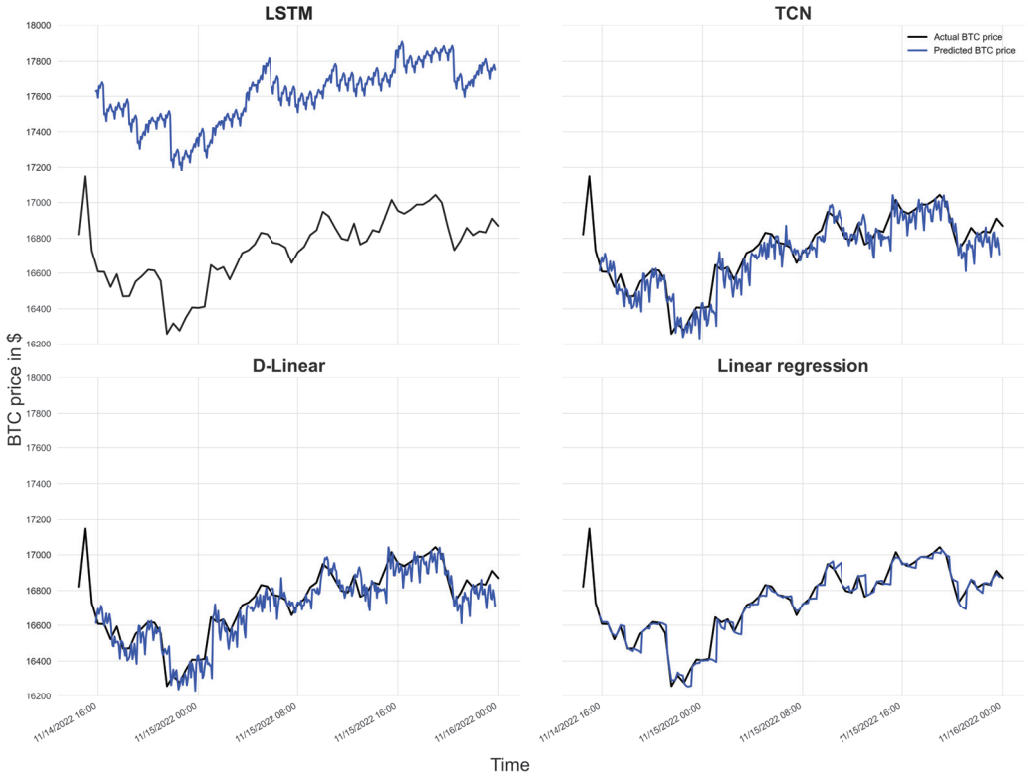


Figure 3. Performance of our models on the last segment of the test set using *mean weighting* and *all sentiment scores* of BERTweet SA.

4.2. Feature Comparison

Next, we compare our different feature setups with each other using only the price, using the price and VADER, using the price and BERTWeet SA (all sentiment scores vs. only the compound score), and using the price and BERTWeet ED using all scores. The results are shown in Table 3.

First, the combination of the BERTweet SA compound score and an LR model yields the highest overall performance. However, BERTweet SA with the compound score alone is not always the best combination of features. Second, for both VADER and BERTweet SA, leveraging the compound score only instead of all the features performed generally better. This seems surprising, especially for the more sophisticated models, which should have the capability to capture what is meaningful and what is not. However, this possibility may have been impeded by the temporal gaps in our data. In addition, BERTweet SA works slightly better than VADER SA. This is exemplified by our best model, which relies solely on the BERTweet SA compound score. However, which one is better depends heavily on the specific model and feature combination. Another observation we make is that BERTweet ED works decently. In most cases, it works better than providing all SA scores (for both VADER and BERTweet) but worse than providing only the compound scores. In general, we find that sentiment scores provide assistance only in specific situations, as is evident from a comparison between Tables 2 and 3. For the more complex models, it even hurts their average performance. We suspect that this may be due to not having enough training data to generalize to these unseen domains and, especially, price ranges.

4.3. Weighting Comparison

Next, we compare the mean weighting scheme (as used above) with our user influence weighting scheme (see Section 3.3.3). The results are shown in Table 4.

Again, we want to highlight a couple of observations. First, except for VADER SA, our novel user influence weighting scheme aids in every feature/model combination (except for LSTM + BERTweet ED, which is an outlier). Sometimes it increases predictive performance by a large margin (e.g., greater than 2$\times$ improvement for TCN + BERTWeet SA). This underlines the merits of our method. However, for VADER SA, our user influence weighing scheme slightly decreases the predictive performance for all but one model (D-Linear), but only insignificantly. Therefore, we conclude that our method encourages further experiments using our user influence weighting, and, additionally, the exploration of other weighting schemes.

4.4. Comparison with Other Research Works

Additionally, we compare our hybrid BTC price forecasting algorithm with other, similar research works in terms of experimental setup and design. However, comparing our research work directly with others is not possible since we employ a unique forecasting scheme and dataset.

To predict the next 4 h price of BTC, we use a combination of sentiment scores derived from microblogs and historical BTC price data from the previous 8 h for each prediction. Serafini et al. [13] utilized an ARIMAX and an LSTM model to predict the price of BTC of the next day, while Edgari et al. [59] utilized an XGBoost model to classify whether BTC would rise or fall in the next minute. Hence, both comprise highly different periods. Both research works employ a similar approach by first collecting historical BTC price data and tweets for SA, be it by simply using it from *Kaggle* [59] or by using a web scraper [13]. Both used VADER to derive sentiment scores for a single tweet. Edgari et al. [59] used accuracy, the AUC (area under curve), and the F1-score as error metrics and showed that adding sentiment scores as inputs to the model considerably improved the performance with respect to each metric. With added sentiment scores, they were able to predict whether BTC would rise or fall with an accuracy of 90%. Serafini et al. [13] showed that the ARIMAX model outperforms the more complex LSTM model, and, using the ARIMAX model, achieved an MSE of 0.14%. Interestingly, they found that using a lower number of

features produces better forecasts compared to using more features, which coincides with our findings.

However, it must be noted that the datasets used by both methods comprise totally different time periods and price ranges compared to ours—specifically, Serafini et al. [13] used data from April 2017 to October 2019, while Edgari et al. [59] used data from February 2021 to December 2021. Although our hybrid method shows higher error metrics compared to the work of Serafini et al. [13], our evaluation is both more robust and realistic, since it spans more diverse price ranges and market scenarios. Moreover, we use more models for comparison, employ two other SA techniques based on BERTweet, and improve the forecasts with our novel weighting scheme. The period under consideration in the work of Edgari et al. [59] is even shorter and, hence, makes evaluating their method very limiting and only useful for the employed 1-min period. In addition, we also employ proper baselines, which both research works we compare with lack [13,59]. Since our work does not feature any of the limitations mentioned, we argue that it is more useful compared to previously employed research works. Other approaches, as specified in Section 2, are even more dissimilar to ours, making any comparison unfeasible.

5. Conclusions and Future Work

In this article, we propose a new hybrid method for predicting the price of Bitcoin, using historic prices and affective cues from tweets. Our method integrates time series forecasting techniques with SA and ED based on a BERTweet model. It also adopts a novel weighting scheme to aggregate multiple sentiment scores from a given time period into a single sentiment score, factoring in the importance or influence of *Twitter* users. We investigate our method on a dataset spanning long and diverse price ranges and show that our approach outperforms previous methods in terms of predictive accuracy. Our empirical results indicate that sentiment scores derived from both VADER and BERTweet can help investors make better forecasts, which supports the hypothesis set in Section 1. Based on our findings, we suggest using simple models such as linear regression to predict the price of Bitcoin instead of more complex ones, particularly if the amount of available data is limited. Moreover, we suggest using BERT-based SA techniques and experimenting with weighing the influence of sentiment scores.

Forecasting the price of BTC is a very challenging task because it is a highly volatile asset. To acknowledge this fact, we created a dataset spanning price ranges from very bullish to bearish periods, thereby following a more realistic experimental setup than in previous works. To accurately forecast prices during such periods, a high degree of generalizability is required from the adopted forecasting model. We achieve this by combining the historical price data with sentiment scores that are derived from a BERT model fine-tuned first on tweets and then on sentiment analysis and are weighted by the number of followers of the author of the tweet; additionally, we use a linear regression model that is less prone to overfitting compared to more complex models. Furthermore, we show that it is precisely overfitting that causes more complex models to have issues. We suspect that they could benefit significantly from an even larger and more diverse dataset. It would be interesting to see whether these models then perform better or even outperform simpler models, as has been shown in previous work, but using different setups and especially market conditions. For now, our recommendation is to use simpler models, such as linear regression. In addition to their better generalization capabilities, they are also faster during inference and easier to use.

We have shown that BERTweet-based SA generally outperforms VADER SA in terms of MAE and RMSE. In addition, we are the first to use BERTweet-based ED for this task. Although it generally did not perform as well as the two investigated SA models, we have shown that it can be used to improve price forecasts, particularly for the more complex models. The more granular classification in ED also adds an additional dimension of interpretability. In future work, it may be of interest to combine these emotion scores with sentiment scores. Finally, it should be noted that including sentiment scores in the

forecasting process often does not improve the forecast accuracy. However, in the cases where it does, it can significantly boost performance.

Moreover, we have shown that weighting the influence of the sentiment scores of a given tweet in a more sophisticated way enables generally better forecasts. In future work, it may be of interest to use other weighting functions, other weighting features (i.e., not just the number of followers but also the number of friends or a combination of the two), or to combine different weighting schemes, either by a simple linear combination or by feeding each of them as input into the model. One limitation of our work is the period under consideration; while being longer than the period in previous research works, it would be of interest to assess model generalizability during even longer periods. Another extension of our work is to study the approach for predicting other cryptocurrencies such as Ethereum. We leave this for future research.

Author Contributions: Conceptualization: M.F., M.K., S.K. and R.W.; data curation: M.F.; formal analysis: M.F. and S.K.; funding acquisition: M.S.; investigation: M.F. and M.K.; methodology: M.F. and S.K.; project administration: M.S.; resources: M.F.; software: M.F.; supervision: M.S.; validation: M.F.; visualization: M.F.; writing—original draft: M.F., M.K., S.K. and R.W.; writing—review and editing: M.F., M.K., S.K., R.W. and M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received support from the Austrian Science Fund (FWF): DFH-23 and from the State of Upper Austria and the Federal Ministry of Education, Science, and Research, through grant LIT-2020-9-SEE-113.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of the data used. The data were obtained from *Kaggle* as well as *Binance* and are available using the respective APIs at [80] and at [79] with the permission of *Kaggle* and *Binance*, respectively.

Acknowledgments: The authors would like to thank Marta Moscati for her invaluable feedback on the manuscript and Kaushik Suresh for publicly releasing the used dataset comprising tweets related to BTC on *Kaggle*.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area Under Curve
BERT	Bidirectional Encoder Representations from Transformers
BTC	Bitcoin
ED	Emotion Detection
ES	Exponential Smoothing
FFT	Fast Fourier Transform
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
LR	Linear regression
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
NLP	Natural Language Processing
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SA	Sentiment Analysis
TCN	Temporal Convolutional Network
VADER	Valence Aware Dictionary and sEntiment Reasoner

References

1. Nakamoto, S. Bitcoin: A Peer-to-Peer Electronic Cash System. 2008. Available online: https://bitcoin.org/bitcoin.pdf (accessed on 1 December 2022).
2. Patel, M.M.; Tanwar, S.; Gupta, R.; Kumar, N. A Deep Learning-based Cryptocurrency Price Prediction Scheme for Financial Institutions. *J. Inf. Secur. Appl.* **2020**, *55*, 102583. [CrossRef]
3. Peterson, T. To the moon: A history of Bitcoin price manipulation. *J. Forensic Investig. Account.* **2021**, *13*, 2. [CrossRef]
4. Sovbetov, Y. Factors influencing cryptocurrency prices: Evidence from bitcoin, ethereum, dash, litcoin, and monero. *J. Econ. Financ. Anal.* **2018**, *2*, 1–27.
5. Schilling, L.; Uhlig, H. Some simple bitcoin economics. *J. Monet. Econ.* **2019**, *106*, 16–26. [CrossRef]
6. Karau, S. Monetary Policy and Bitcoin. *J. Int. Money Financ.* **2023**, 137, 102880. [CrossRef]
7. Vujičić, D.; Jagodić, D.; Ranđić, S. Blockchain technology, bitcoin and ethereum: A brief overview. In Proceedings of the 2018 17th International Symposium INFOTEH-JAHORINA (INFOTEH), East Sarajevo, Bosnia and Herzegovina, 21–23 March 2018; p. 6. [CrossRef]
8. Ciaian, P.; Rajcaniova, M.; d'Artis Kancs. The economics of BitCoin price formation. *Appl. Econ.* **2016**, *48*, 1799–1815. [CrossRef]
9. Mudassir, M.; Bennbaia, S.; Unal, D. *Time-Series Forecasting of Bitcoin Prices Using High-Dimensional Features: A ML Approach*; Springer: Berlin/Heidelberg, Germany, 2020; p. 15. [CrossRef]
10. Pano, T.; Kashef, R. A Complete VADER-Based Sentiment Analysis of Bitcoin (BTC) Tweets during the Era of COVID-19. *Big Data Cogn. Comput.* **2020**, *4*, 33. [CrossRef]
11. Vella Critien, J.; Gatt, A.; Ellul, J. Bitcoin price change and trend prediction through twitter sentiment and data volume. *J. Financ. Innov.* **2022**, *8*, 45. [CrossRef]
12. Pant, D.R.; Neupane, P.; Poudel, A.; Pokhrel, A.K.; Lama, B.K. Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis. In Proceedings of the 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), Kathmandu, Nepal, 25–27 October 2018; pp. 128–132. [CrossRef]
13. Serafini, G.; Yi, P.; Zhang, Q.; Brambilla, M.; Wang, J.; Hu, Y.; Li, B. Sentiment-Driven Price Prediction of the Bitcoin based on Statistical and Deep Learning Approaches. In Proceedings of the 2020 International Joint Conference on Neural Networks, IJCNN, Glasgow, UK, 19–24 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–8. [CrossRef]
14. Wołk, K. Advanced social media sentiment analysis for short-term cryptocurrency price prediction. *Expert Syst.* **2020**, *37*, e12493. [CrossRef]
15. Vyas, P.; Vyas, G.; Dhiman, G. RUemo—The Classification Framework for Russia-Ukraine War-Related Societal Emotions on Twitter through Machine Learning. *Algorithms* **2023**, *16*, 69. [CrossRef]
16. Zhou, B.; Zou, L.; Mostafavi, A.; Lin, B.; Yang, M.; Gharaibeh, N.; Cai, H.; Abedin, J.; Mandal, D. VictimFinder: Harvesting rescue requests in disaster response from social media with BERT. *Comput. Environ. Urban Syst.* **2022**, *95*, 101824. . [CrossRef]
17. Mateen, M. Regulation in the Cryptocurrency Industry. Ph.D. Thesis, University of Missouri–Kansas City, Kansas City, MO, USA, 2023. Available online: https://mospace.umsystem.edu/xmlui/handle/10355/95309 (accessed on 18 April 2023).
18. Fu, S.; Wang, Q.; Yu, J.; Chen, S. FTX collapse: A Ponzi story. *arXiv* **2022**, arXiv:2212.09436
19. Boutsoukis, A. Near Real-Time Cryptocurrency Sentiment Analysis. 2023. Available online: https://repository.ihu.edu.gr/xmlui/bitstream/handle/11544/30143/a.boutsoukis_ds.pdf (accessed on 12 January 2023).
20. Akyildirim, E.; Aysan, A.F.; Cepni, O.; Darendeli, S.P.C. Do investor sentiments drive cryptocurrency prices? *Econ. Lett.* **2021**, *206*, 109980. [CrossRef]
21. Kim, H.J.; Hong, J.S.; Hwang, H.C.; Kim, S.M.; Han, D.H. Comparison of Psychological Status and Investment Style between Bitcoin Investors and Share Investors. *Front. Psychol.* **2020**, *11*, 502295. [CrossRef] [PubMed]
22. Das, N.; Sadhukhan, B.; Chatterjee, T.; Chakrabarti, S. Effect of public sentiment on stock market movement prediction during the COVID-19 outbreak. *Soc. Netw. Anal. Min.* **2022**, *12*, 92. [CrossRef] [PubMed]
23. Gurdgiev, C.; O'Loughlin, D. Herding and anchoring in cryptocurrency markets: Investor reaction to fear and uncertainty. *J. Behav. Exp. Financ.* **2020**, *25*, 100271. [CrossRef]
24. Nandwani, P.; Verma, R. A review on sentiment analysis and emotion detection from text. *Soc. Netw. Anal. Min.* **2021**, *11*, 81. [CrossRef] [PubMed]
25. Barberis, N.; Shleifer, A.; Vishny, R. A Model of Investor Sentiment. *J. Financ. Econ.* **1998**, *49*, 307–343. [CrossRef]
26. Hutto, C.; Gilbert, E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proc. Int. AAAI Conf. Web Soc. Media* **2014**, *8*, 216–225. [CrossRef]
27. Al-Shabi, M. Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining. *IJCSNS* **2020**, *20*, 1.
28. Baly, R.; El-Khoury, G.; Moukalled, R.; Aoun, R.; Hajj, H.; Shaban, K.B.; El-Hajj, W. Comparative Evaluation of Sentiment Analysis Methods Across Arabic Dialects. *Procedia Comput. Sci.* **2017**, *117*, 266–273. [CrossRef]
29. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019*; (Long and Short Papers); Burstein, J., Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; Volume 1, pp. 4171–4186. [CrossRef]

30. Nguyen, D.Q.; Vu, T.; Nguyen, A.T. BERTweet: A pre-trained language model for English Tweets. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 9–14.
31. Liu, Z.; Zhu, Z.; Gao, J.; Xu, C. Forecast Methods for Time Series Data: A Survey. *IEEE Access* **2021**, *9*, 91896–91912. [CrossRef]
32. Lindemann, B.; Müller, T.; Vietz, H.; Jazdi, N.; Weyrich, M. A survey on long short-term memory networks for time series prediction. *Procedia Cirp* **2021**, *99*, 650–655. [CrossRef]
33. Gers, F.A.; Eck, D.; Schmidhuber, J. Applying LSTM to Time Series Predictable through Time-Window Approaches. In *Proceedings of the Artificial Neural Networks—ICANN 2001, Vienna, Austria, 21–25 August 2001*; Dorffner, G., Bischof, H., Hornik, K., Eds.; Springer: Berlin/Heidelberg, Germany, 2001; Volume 2130, pp. 669–676. [CrossRef]
34. Jaquart, P.; Köpke, S.; Weinhardt, C. Machine learning for cryptocurrency market prediction and trading. *J. Financ. Data Sci.* **2022**, *8*, 331–352. [CrossRef]
35. Uras, N.; Marchesi, L.; Marchesi, M.; Tonelli, R. Forecasting Bitcoin closing price series using linear regression and neural networks models. *Peerj Comput. Sci.* **2020**, *6*, e279. [CrossRef] [PubMed]
36. Hamayel, M.J.; Owda, A.Y. A Novel Cryptocurrency Price Prediction Model Using GRU, LSTM and bi-LSTM Machine Learning Algorithms. *AI* **2021**, *2*, 477–496. [CrossRef]
37. Dimitriadou, A.; Gregoriou, A. Predicting Bitcoin Prices Using Machine Learning. *Entropy* **2023**, *25*, 777. [CrossRef] [PubMed]
38. Birjali, M.; Kasri, M.; Hssane, A.B. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowl. Based Syst.* **2021**, *226*, 107134. [CrossRef]
39. Balci, S.; Demirci, G.M.; Demirhan, H.; Sarp, S. Sentiment Analysis Using State of the Art Machine Learning Techniques. In *Proceedings of the Digital Interaction and Machine Intelligence—Proceedings of MIDI'2021 - 9th Machine Intelligence and Digital Interaction Conference, Warsaw, Poland, 9–10 December 2021*; Biele, C., Kacprzyk, J., Kopec, W., Owsinski, J.W., Romanowski, A., Sikorski, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2022; Volume 440, pp. 34–42. [CrossRef]
40. Vyas, P.; Reisslein, M.; Rimal, B.P.; Vyas, G.; Basyal, G.P.; Muzumdar, P. Automated Classification of Societal Sentiments on Twitter With Machine Learning. *IEEE Trans. Technol. Soc.* **2022**, *3*, 100–110. [CrossRef]
41. Hoque, M.U.; Lee, K.; Beyer, J.L.; Curran, S.R.; Gonser, K.S.; Lam, N.S.N.; Mihunov, V.V.; Wang, K. Analyzing Tweeting Patterns and Public Engagement on Twitter During the Recognition Period of the COVID-19 Pandemic: A Study of Two U.S. States. *IEEE Access* **2022**, *10*, 72879–72894. [CrossRef]
42. Aslam, N.; Xia, K.; Rustam, F.; Hameed, A.; Ashraf, I. Using Aspect-Level Sentiments for Calling App Recommendation with Hybrid Deep-Learning Models. *Appl. Sci.* **2022**, *12*, 8522. [CrossRef]
43. Balaji, P.; Haritha, D. An Ensemble Multi-Layered Sentiment Analysis Model (EMLSA) for Classifying the Complex Datasets. *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*. [CrossRef]
44. Loria, S. textblob Documentation. *Release 0.15* **2018**, *2*, 269.
45. Gujjar, J.P.; Kumar, H.P. Sentiment analysis: Textblob for decision making. *Int. J. Sci. Res. Eng. Trends* **2021**, *7*, 1097–1099.
46. Aljedaani, W.; Rustam, F.; Mkaouer, M.W.; Ghallab, A.; Rupapara, V.; Washington, P.B.; Lee, E.; Ashraf, I. Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry. *Knowl.-Based Syst.* **2022**, *255*, 109780. [CrossRef]
47. Aslam, N.; Rustam, F.; Lee, E.; Washington, P.B.; Ashraf, I. Sentiment Analysis and Emotion Detection on Cryptocurrency Related Tweets Using Ensemble LSTM-GRU Model. *IEEE Access* **2022**, *10*, 39313–39324. [CrossRef]
48. Zhang, L.; Wang, S.; Liu, B. Deep learning for sentiment analysis: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1253. [CrossRef]
49. Dos Santos, C.; Gatti, M. Deep convolutional neural networks for sentiment analysis of short texts. In Proceedings of the COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, 23–29 August 2014; pp. 69–78.
50. Zou, Y.; Gui, T.; Zhang, Q.; Huang, X.J. A lexicon-based supervised attention model for neural sentiment analysis. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 21–25 August 2018; pp. 868–877.
51. Pérez, J.M.; Giudici, J.C.; Luque, F. Pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks. *arXiv* **2021**, arXiv:2106.09462.
52. Swathi, T.; Kasiviswanath, N.; Rao, A. An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis. *Appl. Intell.* **2022**, *52*, 13675–13688. [CrossRef]
53. Mehtab, S.; Sen, J. A Robust Predictive Model for Stock Price Prediction Using Deep Learning and Natural Language Processing. *arXiv* **2019**, arXiv:1912.07700. [CrossRef]
54. Ye, Z.; Wu, Y.; Chen, H.; Pan, Y.; Jiang, Q. A Stacking Ensemble Deep Learning Model for Bitcoin Price Prediction Using Twitter Comments on Bitcoin. *Mathematics* **2022**, *10*, 1307. [CrossRef]
55. Sattarov, O.; Jeon, H.S.; Oh, R.; Lee, J.D. Forecasting Bitcoin Price Fluctuation by Twitter Sentiment Analysis. In Proceedings of the 2020 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 4–6 November 2020; pp. 1–4. [CrossRef]
56. Oikonomopoulos, S.; Tzafilkou, K.; Karapiperis, D.; Verykios, V.S. Cryptocurrency Price Prediction using Social Media Sentiment Analysis. In *Proceedings of the 13th International Conference on Information, Intelligence, Systems & Applications, IISA 2022, Corfu, Greece, 18–20 July 2022*; Bourbakis, N.G., Tsihrintzis, G.A., Virvou, M., Eds.; IEEE: Piscataway, NJ, USA, 2022; pp. 1–8. [CrossRef]

57. Abraham, J.; Higdon, D.W.; Nelson, J.; Ibarra, J. Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis. *Smu Data Sci. Rev.* **2018**, *1*, 1.
58. Mittal, A.; Dhiman, V.; Singh, A.; Prakash, C. Short-Term Bitcoin Price Fluctuation Prediction Using Social Media and Web Search Data. In Proceedings of the 2019 Twelfth International Conference on Contemporary Computing (IC3), Noida, India, 8–10 August 2019; pp. 1–6. [CrossRef]
59. Edgari, E.; Thiojaya, J.; Qomariyah, N.N. The Impact of Twitter Sentiment Analysis on Bitcoin Price during COVID-19 with XGBoost. In Proceedings of the 2022 5th International Conference on Computing and Informatics (ICCI), New Cairo, Cairo, Egypt, 9–10 March 2022; pp. 337–342. [CrossRef]
60. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
61. Lea, C.; Vidal, R.; Reiter, A.; Hager, G.D. Temporal Convolutional Networks: A Unified Approach to Action Segmentation. In Proceedings of the Computer Vision–ECCV 2016 Workshops, Amsterdam, The Netherlands, 8–10 and 15–16 October 2016.
62. Zeng, A.; Chen, M.; Zhang, L.; Xu, Q. Are Transformers Effective for Time Series Forecasting? *arXiv* **2022**, arXiv:2205.13504.
63. Herzen, J.; Lässig, F.; Piazzetta, S.G.; Neuer, T.; Tafti, L.; Raille, G.; Pottelbergh, T.V.; Pasieka, M.; Skrodzki, A.; Huguenin, N.; et al. Darts: User-Friendly Modern Machine Learning for Time Series. *J. Mach. Learn. Res.* **2022**, *23*, 1–6.
64. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic Differentiation in Pytorch. 2017. Available online: https://openreview.net/pdf?id=BJJsrmfCZ (accessed on 9 April 2023).
65. Ferdiansyah, F.; Othman, S.H.; Zahilah Raja Md Radzi, R.; Stiawan, D.; Sazaki, Y.; Ependi, U. A LSTM-Method for Bitcoin Price Prediction: A Case Study Yahoo Finance Stock Market. In Proceedings of the 2019 International Conference on Electrical Engineering and Computer Science (ICECOS), Batam, Indonesia, 2–3 October 2019; pp. 206–210. [CrossRef]
66. Guo, H.; Zhang, D.; Liu, S.; Wang, L.; Ding, Y. Bitcoin price forecasting: A perspective of underlying blockchain transactions. *Decis. Support Syst.* **2021**, *151*, 113650. [CrossRef]
67. Sharma, A.; Bhuriya, D.; Singh, U. Survey of stock market prediction using machine learning approach. In Proceedings of the 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 20–22 April 2017; Volume 2, pp. 506–509. [CrossRef]
68. Awan, M.J.; Rahim, M.S.M.; Nobanee, H.; Munawar, A.; Yasin, A.; Zain, A.M. Social Media and Stock Market Prediction: A Big Data Approach. *Comput. Mater. Contin.* **2021**, *67*, 2569–2583. [CrossRef]
69. Jr., E.S.G.; Acar, Y. Fitting the damped trend method of exponential smoothing. *J. Oper. Res. Soc.* **2019**, *70*, 926–930. [CrossRef]
70. Chatfield, C. The Holt-Winters Forecasting Procedure. *J. R. Stat. Soc. Ser. (Appl. Stat.)* **1978**, *27*, 264–279. [CrossRef]
71. Musbah, H.; El-Hawary, M.; Aly, H. Identifying Seasonality in Time Series by Applying Fast Fourier Transform. In Proceedings of the 2019 IEEE Electrical Power and Energy Conference (EPEC), Montreal, QC, Canada, 16–18 October 2019; pp. 1–4. [CrossRef]
72. Brigham, E.O.; Morrow, R.E. The fast Fourier transform. *IEEE Spectr.* **1967**, *4*, 63–70. [CrossRef]
73. Bai, S.; Kolter, J.Z.; Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv* **2018**, arXiv:1803.01271.
74. Schneider, A.; Hommel, G.; Blettner, M. Linear Regression Analysis Part 14 of a Series on Evaluation of Scientific Publications. *Dtsch. Ärzteblatt Int.* **2010**, *107*, 776–82. [CrossRef]
75. Ekaputri, A.P.; Akbar, S. Financial News Sentiment Analysis using Modified VADER for Stock Price Prediction. In Proceedings of the 2022 9th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), Tokoname, Japan, 28–29 September 2022; pp. 1–6. [CrossRef]
76. Hutto, C. GitHub—cjhutto/vaderSentiment: VADER Sentiment Analysis. Available online: https://github.com/cjhutto/vaderSentiment (accessed on 24 November 2022).
77. Nakov, P.; Ritter, A.; Rosenthal, S.; Sebastiani, F.; Stoyanov, V. SemEval-2016 task 4: Sentiment analysis in Twitter. *arXiv* **2019**, arXiv:1912.01973.
78. del Arco, F.M.P.; Strapparava, C.; Lopez, L.A.U.; Martín-Valdivia, M.T. EmoEvent: A multilingual emotion corpus based on different events. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 1492–1498.
79. Binance. Binance API. 2022. Available online: https://www.binance.com/en/binance-api (accessed on 15 October 2022).
80. Kash. Bitcoin Tweets. 2021. Available online: https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweets (accessed on 1 November 2022).
81. Lopez, F. Language Detection Library in Python. 2017. Available online: https://github.com/fedelopez77/langdetect (accessed on 20 November 2022).
82. Robertson, S.E.; Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* **2009**, *3*, 333–389. [CrossRef]
83. Muralidhar, N.; Muthiah, S.; Butler, P.; Jain, M.; Yu, Y.; Burne, K.; Li, W.; Jones, D.; Arunachalam, P.; McCormick, H.S.; et al. Using AntiPatterns to avoid MLOps Mistakes. *arXiv* **2021**, arXiv:2107.00079.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

International Journal of
Financial Studies

MDPI

Article

Enhancing Financial Fraud Detection through Addressing Class Imbalance Using Hybrid SMOTE-GAN Techniques

Patience Chew Yee Cheah, Yue Yang and Boon Giin Lee *

School of Computer Science, University of Nottingham Ningbo China, Ningbo 315100, China;
scypc3@nottingham.edu.cn (P.C.Y.C.); yue.yang2@nottingham.edu.cn (Y.Y.)
* Correspondence: boon-giin.lee@nottingham.edu.cn

Abstract: The class imbalance problem in finance fraud datasets often leads to biased prediction towards the nonfraud class, resulting in poor performance in the fraud class. This study explores the effects of utilizing the Synthetic Minority Oversampling TEchnique (SMOTE), a Generative Adversarial Network (GAN), and their combinations to address the class imbalance issue. Their effectiveness was evaluated using a Feed-forward Neural Network (FNN), Convolutional Neural Network (CNN), and their hybrid (FNN+CNN). This study found that regardless of the data generation techniques applied, the classifier's hyperparameters can affect classification performance. The comparisons of various data generation techniques demonstrated the effectiveness of the hybrid SMOTE and GAN, including SMOTified-GAN, SMOTE+GAN, and GANified-SMOTE, compared with SMOTE and GAN. The SMOTified-GAN and the proposed GANified-SMOTE were able to perform equally well across different amounts of generated fraud samples.

Keywords: class imbalance; data generation; deep learning; financial fraud detection

1. Introduction

The financial sector faces a significant challenge in the form of financial fraud, encompassing various forms of criminal deception aimed at securing financial gains, including activities like telecommunication fraud and credit card skimming. The proliferation of electronic payment technology has propelled online transactions into the mainstream, thereby amplifying the occurrence of fraudulent schemes. The prevalence of these fraudulent transactions has led to substantial losses for financial institutions. However, the large daily transactions pose a challenge for humans in manually identifying fraud. Recently, deep learning techniques have been explored and have shown promising results in detecting financial fraud Alarfaj et al. (2022); Fang et al. (2021); Kim et al. (2019). Unfortunately, most real-world financial fraud datasets suffer from a severe class imbalance issue, where the fraud data's proportion is significantly lower than that of nonfraud. In binary classification, class imbalance often leads to biased predictions favoring the majority class Johnson and Khoshgoftaar (2019). Consequently, the classifier's performance on the minority class is compromised, especially when encountering dissimilar frauds. Overcoming this problem poses a significant challenge, as classifiers are expected to achieve high precision and recall in fraudulent class.

To address this problem, several oversampling methods have been employed to generate minority samples. Synthetic Minority Oversampling TEchnique (SMOTE) interpolates between the existing minority data to synthesize minority samples Chawla et al. (2002). Generative Adversarial Networks (GANs) comprise a discriminator that aims to differentiate between real and generated samples and a generator that strives to deceive the discriminator by synthesizing realistic samples Goodfellow et al. (2014). GANs have shown superior results compared with SMOTE Fiore et al. (2019). However, SMOTE may cause overgeneralization issues. GAN, primarily designed for image generation, is not ideal

for handling the class imbalance problem. To overcome these limitations, SMOTified-GAN employs SMOTE-generated samples instead of random noises as input to the GAN Sharma et al. (2022).

In addition to the aforementioned data generation techniques, other hybrids of SMOTE and GAN are worth exploring. This study presents the following contributions:

1. Introducing two data generation techniques, SMOTE+GAN and GANified-SMOTE, designed to effectively address the class imbalance issue in finance fraud detection.
2. Conducting a comprehensive comparison between the proposed oversampling methods and existing data generation techniques, utilizing precision, recall, and F1-score as key performance metrics.
3. Evaluating the performance of the data generation techniques across various neural network architectures, including a Feed-forward Neural Network (FNN), Convolutional Neural Network (CNN), and the proposed hybrid FNN+CNN.
4. Analyzing the impact of training classifiers on different proportions of the generated minority samples.

2. Related Work

The task of detecting financial fraud can be approached as a binary classification challenge, where classifiers examine the patterns within fraudulent and legitimate transactions to classify new transactions accurately. Consequently, it is crucial to possess an ample and diverse dataset to enable classifiers to grasp the inherent patterns of both transaction categories. Addressing the issue of inadequate fraudulent samples in the training dataset, various methodologies have been introduced to create artificial fraud instances and supplement the original data. These techniques include SMOTE, GAN, and SMOTified-GAN.

SMOTE Chawla et al. (2002) has been widely applied to imbalanced training datasets. More than 85 SMOTE variations were proposed by 2018, including SMOTE+TomekLinks, SMOTE+ENN, Borderline-SMOTE, and Adaptive Synthetic Fernández et al. (2018). Recent studies proposed Radius-SMOTE Pradipta et al. (2021), which prevents overlap among generated samples, and Reduced-Noise SMOTE Arafa et al. (2022), which removes noise after oversampling. In financial fraud detection, SMOTE and its variations have been widely utilized to resample highly imbalanced datasets before training models such as AdaBoost Ileberi et al. (2021) and FNN Fang et al. (2021). Besides the finance domain, SMOTE and its variations have found extensive application in other fields dealing with highly imbalanced datasets. In bio-informatics, SMOTE has been used to discriminate Golgi proteins Tahir et al. (2020) and predict binding hot spots in protein–RNA interactions Zhou et al. (2022). In medical diagnosis, SMOTE and its variations have been employed for diagnosing cervical cancer Abdoh et al. (2018) and prostate cancer Abraham and Nair (2018). SMOTE has also been used to predict diabetes Mirza et al. (2018) and heart failure patients' survival Ishaq et al. (2021).

GANs Goodfellow et al. (2014) and their variations have more recently been employed for generating minority samples to tackle the class imbalance problem. Douzas and Bacao (2018) utilized a conditional GAN (cGAN) which can recover the distribution of training data to generate minority samples. To address the mode collapse issue, Balancing GAN was proposed to generate more diverse and higher-quality minority images Mariani et al. (2018). However, in this technique, the generator and discriminator cannot simultaneously reach their optimal states, leading to the development of IDA-GAN Yang and Zhou (2021). In financial fraud detection, GAN has been employed to generate fraud samples for imbalanced datasets before training classifiers, such as AdaBoost-Decision Tree Mo et al. (2019) and FNN Fiore et al. (2019). These studies have reported that the GAN achieves higher AUC, accuracy, and precision compared with SMOTE. Interestingly, Fiore et al. (2019) found that the best performance was achieved when twice as many GAN-generated fraud samples as the original fraud data were added to the training dataset. In other finance-related domains, GANs have been utilized to address class imbalance in money laundering detection in gambling Charitou et al. (2021). GANs and their variations have also been used

extensively for high-dimensional imbalanced datasets, such as images Mariani et al. (2018); Scott and Plested (2019) and biomedical data Zhang et al. (2018). Recent studies have successfully applied GANs and their variations to generate minority samples in bio-informatics Lan et al. (2020).

Despite the notable accomplishments of SMOTE and GAN, these methods have certain limitations. SMOTE may introduce noise that leads to overgeneralization Bunkhumpornpat et al. (2009). While GANs can generate more "realistic" data, they may not be ideal for handling imbalanced data, as it was originally designed for generating images using random noise. Additionally, there may be insufficient real minority data available for training the GAN Mariani et al. (2018). To address these limitations, Sharma et al. (2022) proposed SMOTifed-GAN, which employs SMOTE-generated samples as input for GAN instead of random numbers, resulting in improved performance compared with SMOTE and GAN.

In early studies, financial fraud detection systems predominantly depended on rule-based methodologies, wherein human expertise in fraud was translated into rules to anticipate fraudulent activities Zhu et al. (2021). However, the evolving behaviors of fraudsters and the increasing size of transaction datasets have posed challenges in identifying fraud-related rules manually. As a result, research has shifted towards machine learning methods, such as naive Bayes, logistic regression, support vector machine, random forest, and decision tree (Ileberi et al. 2021; Ye et al. 2019; Zhu et al. 2021), which can "learn" fraud and nonfraud patterns from given datasets. Nonetheless, machine learning techniques require extensive data preprocessing before training the classifier Alarfaj et al. (2022); Kim et al. (2019); Zhu et al. (2021).

In recent years, deep learning has gained popularity in financial fraud detection due to its superior performance compared with traditional machine learning approaches Alarfaj et al. (2022); Fang et al. (2021); Jurgovsky et al. (2018); Kim et al. (2019). Some studies have approached financial fraud detection as a sequence classification problem, considering the temporal sequence of transactions as a crucial factor. Sequential models, such as Gated Recurrent Units Branco et al. (2020), Long Short-Term Memory (LSTM) Jurgovsky et al. (2018), and Time-aware Attention-based Interactive LSTM Xie et al. (2022), have been proposed. However, since most available financial fraud datasets lack time-sequence information, sequential models may not be suitable in such cases. Due to the vector format of finance fraud datasets without time-sequence information, FNNs are considered a suitable choice Fang et al. (2021); Fiore et al. (2019); Kim et al. (2019). Initially designed for image processing and classification, CNNs have also been found effective in financial fraud detection Alarfaj et al. (2022); Chen and Lai (2021); Zhang et al. (2018). Their 1D convolution layers can extract patterns within smaller segments of a transaction vector.

Building on Fiore et al. (2019)'s findings, this study aimed to assess the performance of a model using varying amounts of minority samples in the training dataset. To achieve this, the study explores the use of SMOTE, GAN, SMOTified-GAN, and other variants of hybrid SMOTE and GAN. Consequently, a combination of SMOTE- and GAN-generated minority samples, along with GANified-SMOTE, was proposed to fulfill the research aims. Finally, FNN, CNN, and FNN+CNN models were employed to ensure a fair evaluation of the performances of different data generation techniques.

3. Methodology

3.1. Data Preprocessing

The experiment utilized the Kaggle (2018) credit card fraud dataset, consisting of 284,807 transactions conducted by European credit card holders over two days in September 2013. This dataset comprises 31 numerical features, including Time, Amount, Class, and 28 other unnamed features. The 'Time' feature represents the elapsed time in seconds since the first transaction, while the 'Amount' feature denotes the transaction amount. The 'Class' label indicates fraudulence, utilizing binary values, where 1 and 0 represent fraud and

nonfraud, respectively. Notably, only 492 transactions (0.172%) are classified as fraudulent, resulting in a highly imbalanced distribution.

To facilitate gradient descent convergence and mitigate bias towards features with larger magnitudes, all features except the 'Class' label were rescaled to the range $[0, 1]$ while maintaining the original feature distribution. This rescaling process for a value X in a given feature was transformed into a new value X' (see Equation (1), where X_{min} and X_{max} represent the minimum and maximum values of the feature, respectively) to maintain the original feature distribution.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

Subsequently, the dataset was divided into a training set comprising 80% of the data (227,451 nonfraud and 394 fraud) and a testing set comprising the remaining 20% (56,864 nonfraud and 98 fraud).

3.2. Data Generation Methods

To address the issue of class imbalance, this study explored five data generation techniques: SMOTE, GAN, and their respective combinations.

3.2.1. SMOTE

SMOTE creates synthetic minority samples rather than duplicating existing ones to avoid overfitting. For a specific minority data point x represented as a vector, a vector x_k is randomly chosen from its k-nearest neighbors to generate a new sample x' using Equation (2). In this study, 394 instances of fraudulent data from the training dataset were utilized with the SMOTE technique, employing five nearest neighbors, to generate additional fraud samples, as depicted in Figure 1.

$$x' = x + rand(0, 1) \times (x - x_k) \tag{2}$$

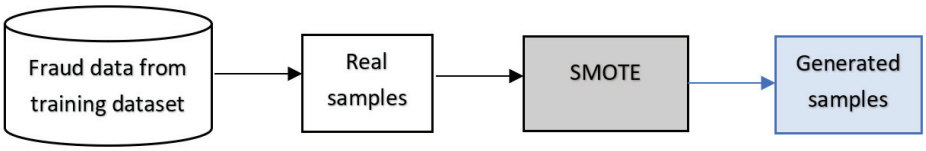


Figure 1. SMOTE employed in this study utilizing five nearest neighbors for random interpolations and generating minority samples.

3.2.2. GAN

A GAN comprises a generator G and a discriminator D that engage in a competitive training process to improve their respective objectives. The discriminator aims to correctly classify real samples x and fake samples generated by the $G(z)$, where z represents random noise or the latent space input to the G. The D's predictions for real and generated samples are denoted as $D(x)$ and $D(G(z))$, respectively. By considering real samples with a label of 1 and generated samples with a label of 0, the D's loss function is defined in Equation (3), where E calculates the error or distance between the D's prediction and the true label. The G's objective is to generate realistic fake samples from random noise that can deceive the D into misclassifying them. The G's general loss function, defined in Equation (4), allows it to improve the quality of the generated samples based on the feedback received from the D's classification. As the G and D continue enhancing their performance, the quality of the generated minority samples improves.

$$L_D = E(D(x), 1) + E(D(G(z)), 0) \tag{3}$$

$$L_G = E(D(G(z)), 1) \tag{4}$$

The proposed GAN architecture, as shown in Figure 2, consists of a 5-layer FNN G with respective neuron counts of 100, 256, 128, 64, and 30. The G takes 100 random noises sampled from a normal distribution. LeakyReLU activation function (Equation (5)) is used in all hidden layers, and dropout layers with a dropout rate of 0.2 are added after each hidden layer to mitigate overfitting. The output layer employs the sigmoid activation function (Equation (6)) to produce values between 0 and 1. Similarly, the D is a 5-layer FNN with identical activation functions and dropout layers. However, the neuron counts are 30, 128, 64, 32, and 1 for each layer. The D employs a stochastic gradient descent (SGD) optimizer with a learning rate of 0.05. The loss function depicted in Figure 2 is binary cross-entropy (Equation (7)), as the D's task involves binary classification. The GAN network also employs an optimizer with the same learning rate as the D, but the loss function utilizes the mean squared error metric (Equation (8), where y_i is the true label and $\hat{y}_i$ is the predicted class) as feedback for the G.

$$f(x) = max(0.1x, x) \tag{5}$$

$$S(x) = \frac{1}{1 + e^{-x}} \tag{6}$$

$$BinaryCrossEntropy = \frac{1}{n}\sum_{i=1}^{n}[y_i \cdot log(\hat{y}_i) + (1 - y_i) \cdot log(1 - \hat{y}_i)] \tag{7}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8}$$

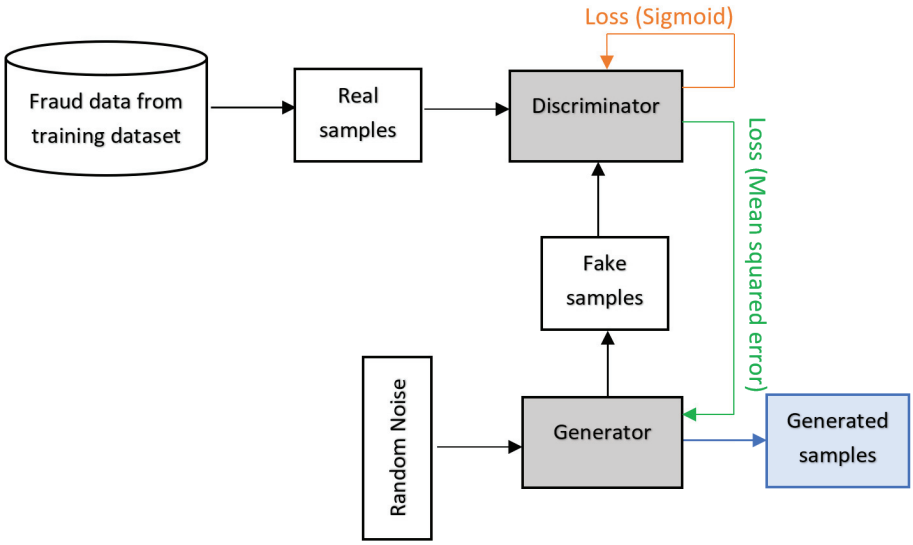


Figure 2. GAN architecture employed in this study consisting of a 5-layer FNN generator and discriminator, leading to the generation of the final minority samples depicted by the blue square.

The fraud data from the training dataset were utilized to train D, enabling it to recognize patterns in real fraud data and generate fraud samples. Since there were only 394 fraudulent data points available for training, the batch size was reduced to 32. The number of training epochs was set to 1000 to allow sufficient time for the G and D to improve their performance. Following training, the G is employed to generate fraud samples based on the required number of minority samples.

3.2.3. SMOTified-GAN

GAN can learn patterns from minority data, resulting in more authentic minority samples. However, using random noise as input for the GAN G can be seen as generating samples from scratch, making it more challenging to train the G to produce high-quality

samples. By utilizing SMOTE-generated samples as input, the generation process becomes simpler as the *G* begins with pre-existing fraud samples (Sharma et al. 2022). In the proposed approach, SMOTE was applied with the five nearest neighbors to generate double the number of fraud samples. Figure 3 illustrates that 788 SMOTE-generated samples were used as input for the GAN *G*. The hyperparameters of the GAN in the SMOTified-GAN model remained the same as the regular GAN, except for the number of neurons in the input layer of the *G*, which was adjusted to 30 to match the 30 features present in the SMOTE-generated fraud samples.

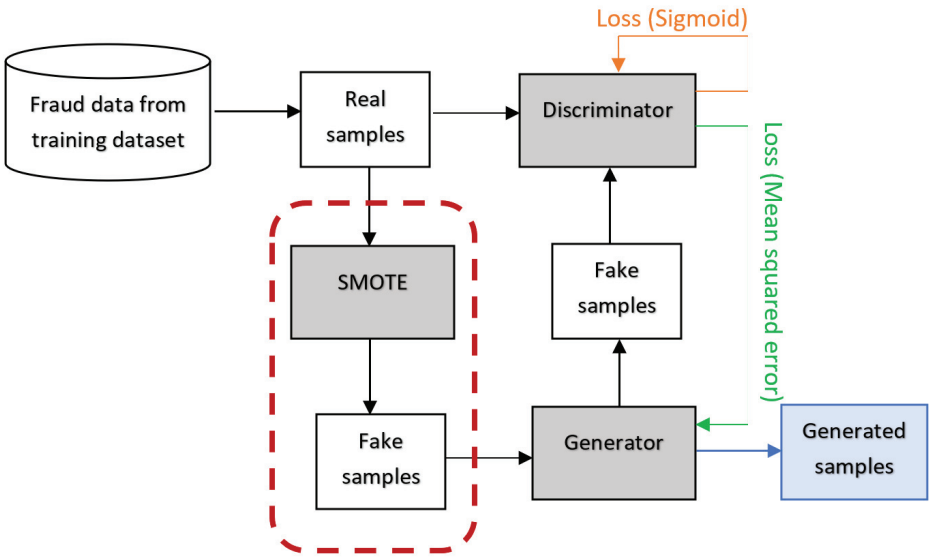


Figure 3. SMOTified-GAN architecture that employed SMOTE-generated samples as the input for the *G*, deviating from the traditional GAN approach that uses random noise. The final minority samples were produced and depicted in the blue square.

3.2.4. SMOTE+GAN

To address the limitations of SMOTE and GAN, a hybrid approach was proposed and employed to enhance the ratio of fraudulent data in the training dataset. The SMOTE-generated and GAN-generated fraud samples were directly combined with the original training dataset without any alterations, as depicted in Figure 4. The combined dataset comprised an equal contribution from both the SMOTE- and GAN-generated samples, amounting to half of the total required generated data.

3.2.5. GANified-SMOTE

Another hybrid method, GANified-SMOTE, was implemented. The random interpolation makes SMOTE-generated samples susceptible to the noise present in the dataset. Consequently, the generated minority samples are located near the boundary of the majority class, leading to higher misclassification rates. Conversely, GAN can learn the underlying patterns of the minority class, reducing the impact of such noise. By utilizing GAN-generated data for SMOTE interpolations, the limitations of SMOTE can be overcome. Additionally, applying SMOTE on the GAN-generated data can decrease reliance on the prominent patterns of the minority class, thereby mitigating overfitting. Figure 5 illustrates the utilization of fraud samples generated by the GAN, which are then processed with SMOTE to generate the necessary number of fraud samples. The resulting output from SMOTE is combined with the original training dataset, which originally contained 394 authentic fraud data points.

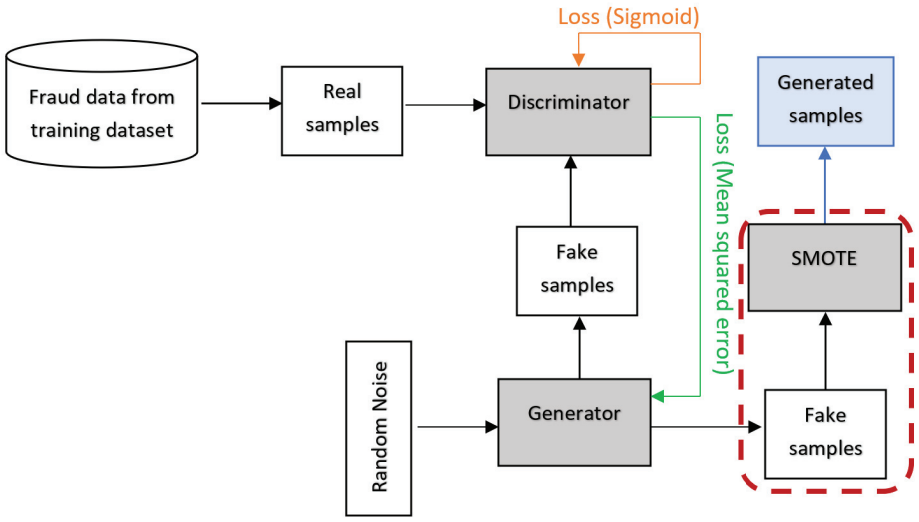


Figure 4. SMOTE+GAN architecture was employed to generate minority samples by directly incorporating SMOTE and GAN techniques. These generated samples were then merged with the original training dataset.

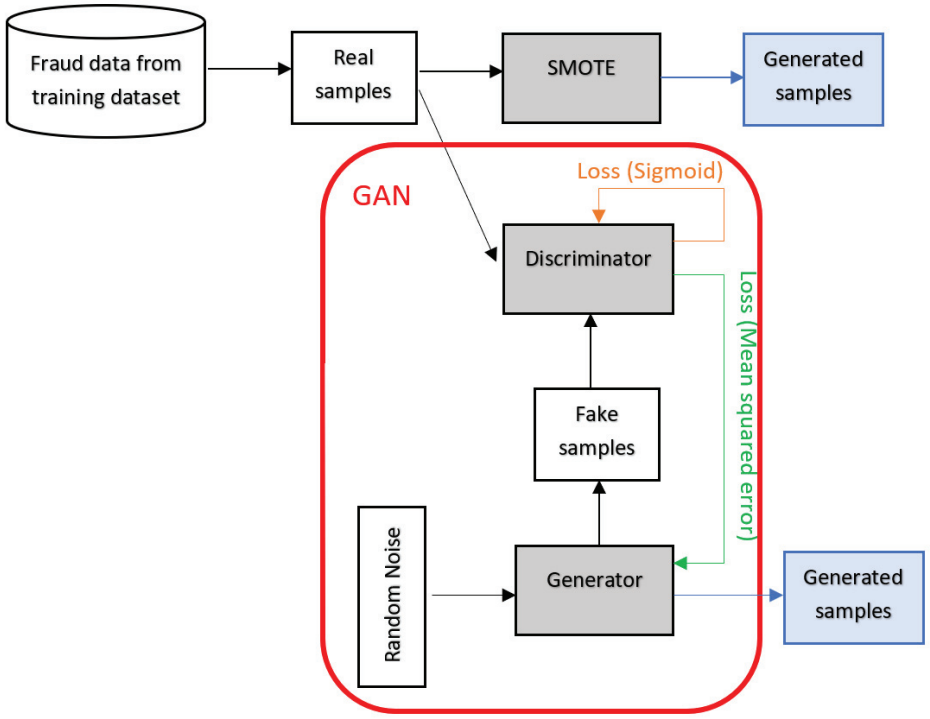


Figure 5. Proposed GANified-SMOTE architecture that involved the application of SMOTE to the GAN-generated samples, as indicated by the red dashed square.

3.3. Summary of Data Generation Methods

Table 1 presents an overview of the types and quantities of fraud data utilized in each data generation method. Two experiments were conducted for each method to assess the impact of varying amounts of generated data in the training dataset. In the first experiment (Test A), the training dataset was adjusted to achieve a balanced distribution of 50% fraud and 50% nonfraud samples. In the second experiment (Test B), only 788 fraud samples were generated, twice the number of the original fraud data in the training dataset. This choice was based on the finding (Fiore et al. 2019) that injecting twice as many GAN-generated fraud samples as the original fraud data produced the optimal outcome.

For both experiments, fraud samples were generated using five data generation techniques after splitting the complete dataset into training and testing sets. The testing dataset was not utilized for data generation to ensure that the validation conducted using these unseen data reflects the model's performance when applied to real-world financial fraud detection systems, as these systems encounter unseen data.

Table 1. Variations in the types and quantities of fraud data utilized for each data generation method.

Generation Method	Types of Fraud Data	Total Fraud Data	Total Generated Fraud Data	
			Test A	Test B
SMOTE	Real	394	227,057	788
GAN	Real	394	227,057	788
SMOTified-GAN	Real SMOTE	394 788	227,057	788
SMOTE+GAN	Real	394	SMOTE: 113,529 GAN: 113,529	SMOTE: 344 GAN: 344
GANified-SMOTE	Real SMOTE	394 394	227,057	788

3.4. Deep Learning Models

3.4.1. FNN

A preliminary investigation was conducted to assess several hyperparameter configurations of the FNN (Feed-forward Neural Network) along with SMOTE-generated samples to tackle the problem of class imbalance. The two most effective models were selected as classifiers to evaluate all the data generation techniques. Table 2 contains the hyperparameters used for these models. Both models employed the Rectified Linear Unit (ReLU) activation function for their hidden layers. To counter overfitting, dropout layers with a dropout rate of 0.1 were inserted after each hidden layer. The output layer utilized the sigmoid activation function to ensure that the output probabilities fall within the range of 0 to 1, representing the likelihood of a transaction being fraudulent. For the loss function, binary cross-entropy was employed. Due to the substantial size of the training dataset, a batch size of 128 was chosen, and the training process was executed over 100 epochs, allowing for multiple iterations to refine the model.

Table 2. The hyperparameters of the two top performing models within FNN and CNN variations.

Model	Layers	Neurons in Each Layer	Filters	Optimizer	Learning Rate
FNN1	6 DENSE	30-200-100-50-10-1	-	SGD	0.01
FNN2	6 DENSE	30-200-100-50-10-1	-	SGD	0.05
CNN1	1 CONV+1 POOL	-	64	Adam	0.01
CNN2	1 CONV+1 POOL	-	32	Adam	0.01

3.4.2. CNN

Similarly to the FNN, various hyperparameter configurations of the CNN were tested, and the two best-performing models were selected for further investigation. The hyperparameters for these models are presented in Table 2. Both models began with an input layer of dimensions (30, 1). Subsequently, a 1D convolutional layer and a max-pooling layer were incorporated, followed by a flattening layer and a dense layer consisting of 50 neurons, utilizing the ReLU activation function, along with a dropout layer featuring a dropout rate of 0.1. The output layer consisted of a single neuron activated by the sigmoid function. The kernel size and pool size for both models were set to 3 and 2, respectively. The initial findings indicated that the CNN models reached a stable loss and accuracy after the 50th

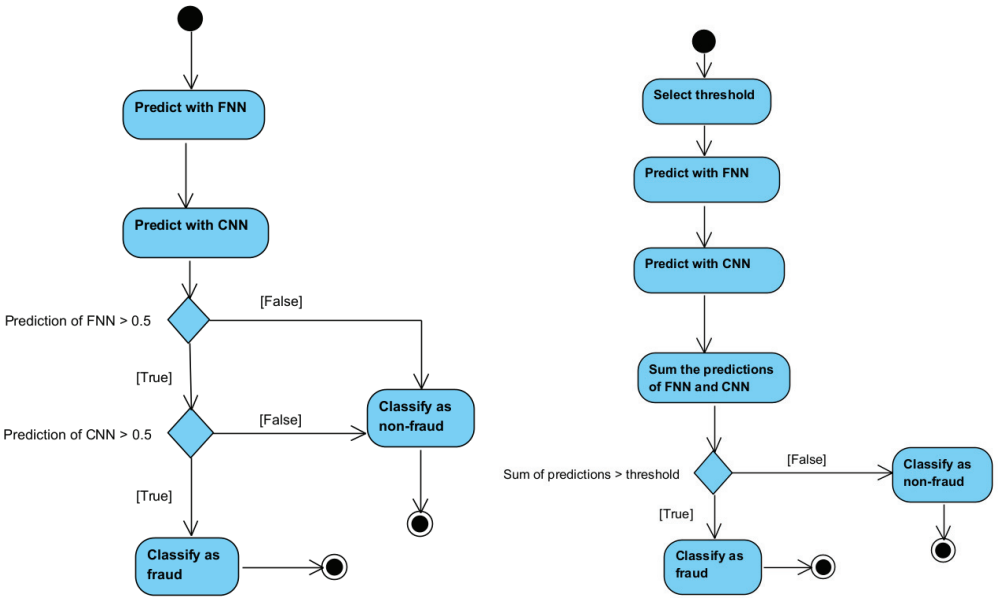
epoch. Consequently, the number of training epochs was set to 50, providing sufficient time to refine the models and observe their performance.

3.4.3. FNN+CNN

FNN and CNN models tend to misclassify nonfraudulent transactions, while demonstrating an intuitive ability to identify the same fraudulent transactions. Consequently, this study integrated the two models to enhance the final prediction, aiming to reduce the false-positive rate within the fraud class. By leveraging the strengths of both models and combining their insights, it was anticipated that the integrated approach would yield improved accuracy and more reliable identification of fraudulent transactions.

Method I: The final prediction is classified as fraudulent only if both FNN and CNN predict the transaction as fraudulent. The detailed processes and decision steps of this method are depicted in Figure 6a. Both the FNN and CNN output a probability of a transaction being fraudulent, where a value greater than 0.5 is considered fraudulent. Hence, the integrated model predicted a transaction as fraudulent only if both models' output surpassed 0.5. Intuitively, it is improbable for a nonfraudulent transaction to be classified as fraudulent by both models, given their tendency to learn distinct patterns associated with fraud and nonfraud. This integration reinforces the reliability of the fraud prediction, since it must satisfy both conditions.

Method II: The initial study demonstrated that the first method successfully enhanced the precision of fraud detection but resulted in a decrease in the recall. Therefore, another method was proposed to increase recall while maintaining a high precision. In certain instances, one of the models produces a value close to 1, indicating a high probability of the transaction being fraudulent. Conversely, the other model generates a value below but close to 0.5. According to the first method, these transactions would be predicted as nonfraudulent. However, intuitively, such transactions are more likely to be fraud, since one of the models strongly indicates fraud. To address this scenario, the sum of the output values from both models is utilized to make the final prediction. If the sum exceeds a selected threshold, the prediction will be fraudulent. The detailed processes and decision steps of this method are depicted in Figure 6b.



(**a**) Method I. (**b**) Method II.

Figure 6. Flowchart illustrating the two methods employed for FNN+CNN.

4. Results and Discussion

4.1. Deep Learning Models with SMOTE-Generated Data

In this study, two FNN and two CNN models were developed to determine the optimal configuration, and their specific hyperparameters are outlined in Table 2. The training process took place on a machine equipped with an 11th Gen Intel Core i7-11375H CPU, 16GB RAM, Intel Iris Xe Graphics, and NVIDIA GeForce RTX 3060 Laptop GPU. The training dataset consisted of an equal distribution of fraud and nonfraud data, with the fraud samples generated using the SMOTE technique (refer to Section 3.2.1). To evaluate their performance, these variations were tested on the testing dataset, and various metrics were employed, including training accuracy, loss, and time, as well as testing precision (PR), recall (RC), F1-score (F1), and root mean squared error (RMSE).

In Table 3, all the top-performing models achieved impeccable precision, recall, and F1-score ($PR = RC = F1 = 1.00$) for the nonfraud class, and their recalls in the fraud class were satisfactory ($RC \geq 0.85$). However, their precision and F1-score in the fraud class did not meet the desired criteria. FNN1, which utilized a lower learning rate, demonstrated higher precision and F1-score compared with FNN2. Similarly, CNN2, with fewer filters, exhibited higher precision and F1-score compared with CNN1, albeit with a slightly lower precision. Consequently, FNN2 and CNN2, boasting the highest F1-score within their respective FNN and CNN models, were selected as the top-performing models for integration (FNN+CNN). It is worth noting that FNN models using the SGD optimizer yielded superior results compared with those employing Adam, while the opposite was observed for CNN. Additionally, CNN's training time was longer than that of FNN due to the fewer epochs utilized.

Table 3. The results of the two top-performing models within FNN and CNN variations are presented. The selection of the best model was based on comparing their recalls first, followed by their precision.

Model	TA [1]	TL [2]	TT [3]	PR [4]		RC [5]		F1 [6]		RMSE [7]
				NF [8]	F [9]	NF [8]	F [9]	NF [8]	F [9]	
FNN1	0.9952	0.0169	27 m 00 s	1.00	0.25	1.00	0.91	1.00	0.39	0.0692
FNN2	0.9961	0.0137	27 m 22 s	1.00	0.29	1.00	0.90	1.00	0.44	0.0624
CNN1	0.9960	0.0196	13 m 00 s	1.00	0.29	1.00	0.89	1.00	0.43	0.0634
CNN2	0.9981	0.0123	12 m 54 s	1.00	0.47	1.00	0.85	1.00	0.60	0.0437

[1] Training Accuracy, [2] Training Loss, [3] Training Time, [4] Precision, [5] Recall, [6] F1-score, [7] Root Mean Square Error, [8] Nonfraud, [9] Fraud.

4.2. Deep Learning Models with Data Generation Methods

4.2.1. FNN and CNN

This research employed four selected FNNs and CNNs to evaluate the impact and performance of the generated data. The results can be found in Table 4. When employing the identical data generation method and incorporating an equal quantity of fraudulent data, both versions of FNN and CNN yielded comparable outcomes. This demonstrates that the outcomes derived from a data generation method are not significantly influenced by the classifier's parameters.

Table 4. The outcomes of employing five methods to generate data on four models were evaluated on two distinct minority samples.

Model	Generation Method	Test Samples	TA [1]	TL [2]	PR [3]	RC [4]	F1 [5]
FNN1	SMOTE	A	0.9952	0.0169	0.25	**0.91**	0.39
		B	0.9994	0.0028	0.81	0.88	0.84
	GAN	A	0.9995	0.0036	0.84	0.88	**0.86**
		B	0.9994	0.0025	0.81	0.88	0.84
	SMOTified-GAN	A	0.9995	0.0024	**0.85**	0.87	**0.86**
		B	0.9995	0.0024	**0.83**	0.86	0.84
	SMOTE+GAN	A	0.9966	0.0119	0.32	0.90	0.48
		B	0.9995	0.0025	**0.83**	0.88	**0.85**
	GANified-SMOTE	A	0.9994	0.0038	0.81	0.87	0.84
		B	0.9995	0.0025	0.82	0.88	**0.85**
FNN2	SMOTE	A	0.9961	0.0137	0.29	0.90	0.44
		B	0.9995	0.0026	0.85	0.88	0.86
	GAN	A	0.9995	0.0034	0.83	0.88	0.85
		B	0.9994	0.0027	0.81	0.88	0.84
	SMOTified-GAN	A	0.9996	0.0023	0.87	0.88	**0.87**
		B	0.9995	0.0024	**0.87**	0.87	**0.87**
	SMOTE+GAN	A	0.9956	0.0141	0.27	**0.91**	0.41
		B	0.9995	0.0025	0.85	0.88	0.86
	GANified-SMOTE	A	0.9996	0.0029	**0.88**	0.86	**0.87**
		B	0.9995	0.0022	**0.87**	0.87	**0.87**
CNN1	SMOTE	A	0.9960	0.0196	0.29	**0.89**	0.43
		B	0.9995	0.0029	0.85	0.87	0.86
	GAN	A	0.9995	0.0042	0.85	0.87	**0.86**
		B	0.9995	0.0031	0.86	0.86	0.86
	SMOTified-GAN	A	0.9995	0.0028	**0.90**	0.83	**0.86**
		B	0.9994	0.0032	0.80	0.84	0.82
	SMOTE+GAN	A	0.9962	0.0140	0.29	0.86	0.44
		B	0.9996	0.0025	**0.87**	**0.88**	**0.87**
	GANified-SMOTE	A	0.9995	0.0045	0.85	0.84	0.85
		B	0.9994	0.0034	0.79	0.86	0.82
CNN2	SMOTE	A	0.9981	0.0123	0.47	0.85	0.60
		B	0.9995	0.0034	0.83	0.86	0.84
	GAN	A	0.9994	0.0040	0.82	0.87	0.84
		B	0.9995	0.0029	0.86	0.84	0.85
	SMOTified-GAN	A	0.9995	0.0026	**0.89**	0.83	**0.86**
		B	0.9994	0.0036	0.81	**0.88**	0.84
	SMOTE+GAN	A	0.9965	0.0181	0.32	**0.90**	0.47
		B	0.9996	0.0039	0.88	0.86	**0.87**
	GANified-SMOTE	A	0.9995	0.0035	**0.89**	0.83	**0.86**
		B	0.9996	0.0023	**0.91**	0.83	**0.87**

[1] Training Accuracy, [2] Training Loss, [3] Precision, [4] Recall, [5] F1-score. The top-performing PR, RC, and F1-score for both Test A and Test B are highlighted in bold, except for the RC in Test B, as identifying the best result posed a challenge due to high similarities.

The FNN generally yielded a higher recall compared with the CNN, albeit with lower precision. In test A, both SMOTE and SMOTE+GAN exhibited significantly lower precision and F1-score across all models, despite demonstrating a high recall. However, there was substantial improvement observed in test B for these two methods, where synthetic fraud samples were injected at a ratio of twice the original records. Additionally, SMOTE+GAN

achieved the highest F1-score in three out of four models during test B. The proposed GANified-SMOTE generally yielded slightly higher precision and F1-score than GAN, despite having a lower recall. This can be attributed to GAN's ability to capture the original fraud data's characteristics, resulting in the GAN-generated fraud samples being clustered in regions with a high concentration of the original fraud data. The SMOTE's application on the GAN-generated samples generates additional fraud samples in between them, potentially causing a 'blurring' effect on the fraud data's features. This could explain the generally lower recall of GANified-SMOTE in comparison with GAN. As a trade-off, GANified-SMOTE achieves a lower false positive rate, leading to higher precision compared with GAN. When compared with SMOTified-GAN, the proposed GANified-SMOTE generally demonstrated slightly lower precision and F1-score, while maintaining a similar recall. SMOTified-GAN generates fraud samples using SMOTE-generated samples as input, which can result in the production of more realistic and diverse samples. The SMOTified-GAN-generated samples tended to be more centrally distributed within fraud areas and less centrally distributed within nonfraud areas, which could explain the higher precision and F1-score observed.

4.2.2. FNN+CNN

The top-performing models from the FNN and CNN were combined to create two distinct hybrid FNN+CNN methods, as illustrated in Figure 6. The results of the FNN+CNN approach for Method I and II are depicted in Table 5. Method I exhibited improved precision for detecting fraudulent cases compared with using FNN or CNN models alone. In Test A, the SMOTE and SMOTE+GAN showed significant improvements in precision, despite a slight decrease in recall, particularly when compared with the FNN model. This decline can be attributed to the fact that fraud predictions must meet two distinct conditions, resulting in a reduced number of predicted fraud cases. Consequently, the fraud class's precision increases, since precision is determined by the ratio of true fraud cases to predicted fraud cases. However, this trade-off leads to a decrease in the fraud class's recall. Nevertheless, the overall F1-score exhibited a slight increase compared with the individual FNN and CNN models. In Method II, different thresholds ranging from 1.1 to 1.9 were tested to determine the optimal threshold value. Since Method II's goal was to enhance recall, the best threshold value was determined based on recall.

Overall, Method II yielded better results than Method I. The findings demonstrated that the proposed hybrid FNN+CNN approach in Method II outperformed the FNN and CNN models individually. Similar to the observations on the FNN and CNN, injecting twice the number of fraud samples as the original fraud data using SMOTE and SMOTE+GAN yielded better performance than a 50:50 distribution of fraud and nonfraud samples. The performance of GAN, SMOTified-GAN, and GANified-SMOTE was not significantly affected by the number of injected fraud samples. The proposed GANified-SMOTE technique achieved the highest precision for both integration methods and also exhibited high F1-score and recall. This may be attributed to the pure variations in the FNN and CNN used in the hybrid model performing well with GANified-SMOTE. However, the GANified-SMOTE's performance on the FNN and CNN variations was similar. Therefore, it can be concluded that the proposed GANified-SMOTE can achieve high performance regardless of the number of injected fraud samples.

Table 5. The integration of FNN2+CNN2 with Method I and II was evaluated using different data generation techniques on two minority samples, and the top-performing results for each measurement in tests A and B are highlighted in bold.

Method	Generation Method	Test Samples	Threshold	PR [1]	RC [2]	F1 [3]
I	SMOTE	A	-	0.69	0.84	0.76
		B	-	0.88	0.86	0.87
	GAN	A	-	0.88	0.87	0.87
		B	-	0.86	0.84	0.85
	SMOTified-GAN	A	-	0.90	0.83	0.86
		B	-	0.88	**0.87**	0.87
	SMOTE+GAN	A	-	0.56	**0.90**	0.69
		B	-	0.89	0.86	**0.88**
	GANified-SMOTE	A	-	**0.93**	0.83	**0.88**
		B	-	**0.91**	0.83	0.87
II	SMOTE	A	1.1	0.51	0.86	0.64
		B	1.3	0.89	0.87	0.88
	GAN	A	1.5	0.88	0.87	0.87
		B	1.1	0.86	0.86	0.86
	SMOTified-GAN	A	1.1	0.88	0.86	0.87
		B	1.3	0.89	0.87	0.88
	SMOTE+GAN	A	1.4	0.57	**0.90**	0.70
		B	1.3	0.89	0.87	0.88
	GANified-SMOTE	A	1.4	**0.94**	0.85	**0.89**
		B	1.3	**0.91**	0.86	0.88

[1] Precision, [2] Recall, [3] F1-score.

4.3. Comparison with Existing Studies

To evaluate the proposed methodologies, a comparison was made with previous studies that utilized the same dataset, as presented in Table 6. Given the trade-off between precision and recall, attaining flawless outcomes for models, whether existing or proposed, remains elusive. The outcomes observed by (Fiore et al. 2019) and (Sharma et al. 2022) upon applying SMOTE, GAN, and SMOTified-GAN exhibited relatively modest recalls (below 0.80), indicating limited detection of fraudulent transactions. Consequently, their F1-scores generally trailed behind those of the proposed methods. This serves to illustrate the proficiency of the proposed models in effectively identifying fraudulent transactions while upholding a minimal misclassification rate for nonfraudulent data.

All the implemented techniques exhibited higher recall rates compared with the existing studies. However, this improvement came at the expense of lower precision when compared with previous research. One potential explanation for this discrepancy could be the differences in the classifiers utilized. Previous studies Fiore et al. (2019); Sharma et al. (2022) employed classifiers with less than four layers, whereas the proposed classifier consisted of at least four. Consequently, the enhanced classifier was able to better learn the distinguishing characteristics of fraudulent data, improving the identification of such instances. However, this also led to an increased misclassification of nonfraudulent data.

Another factor that could have influenced the outcomes is the stochastic nature of the SMOTE (and the GAN) and deep learning models (the FNN). Despite the lower precision, the F1-score of the proposed methodologies surpassed that of the previous studies, except for SMOTE on Test A. This observation highlights the significant impact of classifier parameters on its performance, irrespective of the data generation methods employed. This observation aligns with the previous findings, indicating an overall enhancement in the F1-score when utilizing a hybrid of FNN and CNN, regardless of the specific data generation methods employed.

Table 6. Comparison of results obtained by existing studies and the proposed methods. Test A is the result for including twice-generated fraud samples as much as the original fraud samples, whereas Test B is the result for 50:50 fraud and nonfraud distributions. The highest Precision, Recall, and F1-score are highlighted in bold.

Generation Method	Test Samples	Classifier	PR [a]	RC [b]	F1 [c]
SMOTE (Sharma et al. 2022)	A	FNN	0.80	0.69	0.71
SMOTE	A	FNN	0.29	0.90	0.44
SMOTE (Fiore et al. 2019)	B	FNN	**0.97**	0.69	0.81
SMOTE	B	FNN	0.85	0.88	0.86
GAN (Sharma et al. 2022)	A	FNN	0.84	0.80	0.81
GAN	A	FNN	0.83	0.88	0.85
GAN (Fiore et al. 2019)	B	FNN	0.93	0.73	0.82
GAN	B	FNN	0.81	0.88	0.84
SMOTified-GAN (Sharma et al. 2022)	A	FNN	0.85	0.80	0.81
SMOTified-GAN	A	FNN	0.87	0.88	0.87
SMOTified-GAN	B	FNN	0.87	0.87	0.87
SMOTE+GAN	A	FNN	0.27	**0.91**	0.41
SMOTE+GAN	B	FNN	0.85	0.88	0.86
SMOTE+GAN	A	FNN+CNN (Method II)	0.57	0.90	0.70
SMOTE+GAN	B	FNN+CNN (Method II)	0.89	0.87	0.88
GANified-SMOTE	A	FNN	0.88	0.86	0.87
GANified-SMOTE	B	FNN	0.87	0.87	0.87
GANified-SMOTE	A	FNN+CNN (Method II)	0.94	0.85	**0.89**
GANified-SMOTE	B	FNN+CNN (Method II)	0.91	0.86	0.88

[a] Precision, [b] Recall, [c] F1-score.

Nonetheless, data generation methods can still affect the performance of the same classifier in different variations. The results from Fiore et al. (2019) and Sharma et al. (2022) demonstrated an increase in recall and F1-score when employing GAN as opposed to SMOTE. However, in this study, the implemented GAN did not improve recall but only enhanced the F1-score on Test A. The challenges in training the GAN may have resulted in the lower quality of generated fraudulent samples. Conversely, SMOTE's random interpolation may not effectively capture the distinguishing characteristics of fraud instances. Therefore, combining SMOTE and GAN in a hybrid approach could result in the two complementing each other and better a representation of the fraudulent data. The proposed SMOTE+GAN demonstrated a slight improvement in recall on Test A compared with SMOTE. Additionally, the implemented SMOTified-GAN and the proposed GANified-SMOTE successfully improved the F1-score.

5. Conclusions

The present study introduces SMOTE+GAN and GANified-SMOTE techniques as innovative solutions to counteract class imbalance, thereby offering financial institutions an effective tool for reducing losses due to fraudulent activities. Additionally, the integration of FNN and CNN in predicting transaction categories is proposed. The effectiveness of the newly proposed data generation methods was assessed against existing techniques using an FNN, CNN, and FNN+CNN as classifiers. The outcomes highlight the potency of GANified-SMOTE, particularly when coupled with the proposed FNN+CNN classifier, in augmenting the F1-score for fraudulent data. This high F1-score indicates the method's capacity to identify a substantial portion of fraudulent transactions with reduced misclassification of legitimate transactions. Notably, GANified-SMOTE and SMOTified-GAN consistently exhibit commendable performance across varying quantities of generated minority samples. Furthermore, the research underscores the significant impact of the classifier's hyperparameter settings on classification performance, irrespective of the employed data generation methods.

In light of this experiment utilizing an online-acquired dataset, it is crucial to recognize that the study's findings may not perfectly simulate real-world scenarios marked by ever-evolving fraudulent behaviors. Future endeavors should validate the efficacy of the proposed methods within actual financial institutions. Moreover, while the experiment employs a labeled dataset with presumed accurate class labels, real-world datasets often pose the challenge of being unlabeled and necessitating comprehensive preprocessing. To tackle class labeling issues, future investigations could explore the potential of unsupervised learning in data generation. Furthermore, to firmly establish the effectiveness of the proposed methods, this study acknowledges that comparisons with existing research were limited. Factors like classifier selection may have influenced observed improvements. Therefore, to enhance generalizability, future research should involve additional classifiers and ablation studies. These efforts would serve to validate the performance of the data generation methods in diverse scenarios.

Author Contributions: Conceptualization, P.C.Y.C. and B.G.L.; methodology, P.C.Y.C.; software, P.C.Y.C.; validation, P.C.Y.C. and Y.Y.; formal analysis, P.C.Y.C.; investigation, P.C.Y.C.; resources, P.C.Y.C. and B.G.L.; data curation, P.C.Y.C.; writing—original draft preparation, P.C.Y.C.; writing—review and editing, B.G.L.; visualization, P.C.Y.C.; supervision, B.G.L.; project administration, B.G.L.; funding acquisition, B.G.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ningbo Science and Technology Bureau grant number 2021B-008-C.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in Kaggle-Credit Card Fraud Detection, reference number Kaggle (2018).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
FNN	Feed-forward Neural Network
SMOTE	Synthetic Minority Oversampling TEchnique
GAN	Generative Adversarial Network

References

Abdoh, Sherif F., Mohamed Abo Rizka, and Fahima A. Maghraby. 2018. Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. *IEEE Access* 6: 59475–85. [CrossRef]

Abraham, Bejoy, and Madhu S. Nair. 2018. Computer-aided diagnosis of clinically significant prostate cancer from MRI images using sparse autoencoder and random forest classifier. *Biocybernetics and Biomedical Engineering* 38: 733–44. [CrossRef]

Alarfaj, Fawaz Khaled, Iqra Malik, Hikmat Ullah Khan, Naif Almusallam, Muhammad Ramzan, and Muzamil Ahmed. 2022. Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. *IEEE Access* 10: 39700–15. [CrossRef]

Arafa, Ahmed, Nawal El-Fishawy, Mohammed Badawy, and Marwa Radad. 2022. RN-SMOTE: Reduced noise SMOTE based on DBSCAN for enhancing imbalanced data classification. *Journal of King Saud University—Computer and Information Sciences* 34: 5059–74. [CrossRef]

Branco, Bernardo, Pedro Abreu, Ana Sofia Gomes, Mariana S. C. Almeida, João Tiago Ascensão, and Pedro Bizarro. 2020. Interleaved sequence RNNs for fraud detection. Paper presented at the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20, New York, NY, USA, August 24–27; Melbourne: Association for Computing Machinery, pp. 3101–9. [CrossRef]

Bunkhumpornpat, Chumphol, Krung Sinapiromsaran, and Chidchanok Lursinsap. 2009. Safe-Level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Advances in Knowledge Discovery and Data Mining*. Edited by Thanaruk Theeramunkong, Boonserm Kijsirikul, Nick Cercone and Tu-Bao Ho. Berlin and Heidelberg: Springer, pp. 475–82.

Charitou, Charitos, Simo Dragicevic, and Artur d'Avila Garcez. 2021. Synthetic data generation for fraud detection using GANs. *arXiv* arXiv:2109.12546.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16: 321–57. [CrossRef]

Chen, Joy, and Kong-Long Lai. 2021. Deep convolution neural network model for credit-card fraud detection and alert. *Journal of Artificial Intelligence and Capsule Networks* 3: 101–12. [CrossRef]

Douzas, Georgios, and Fernando Bacao. 2018. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications* 91: 464–71. [CrossRef]

Fang, Weiwei, Xin Li, Ping Zhou, Jingwen Yan, Dazhi Jiang, and Teng Zhou. 2021. Deep learning anti-fraud model for internet loan: Where we are going. *IEEE Access* 9: 9777–84. [CrossRef]

Fernández, Alberto, Salvador Garcia, Francisco Herrera, and Nitesh V. Chawla. 2018. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research* 61: 863–905. [CrossRef]

Fiore, Ugo, Alfredo De Santis, Francesca Perla, Paolo Zanetti, and Francesco Palmieri. 2019. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences* 479: 448–55. [CrossRef]

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems* 27: 2672–80.

Ileberi, Emmanuel, Yanxia Sun, and Zenghui Wang. 2021. Performance evaluation of machine learning methods for credit card fraud detection using SMOTE and AdaBoost. *IEEE Access* 9: 165286–94. [CrossRef]

Ishaq, Abid, Saima Sadiq, Muhammad Umer, Saleem Ullah, Seyedali Mirjalili, Vaibhav Rupapara, and Michele Nappi. 2021. Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. *IEEE Access* 9: 39707–16. [CrossRef]

Johnson, Justin M., and Taghi M. Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data* 6: 27. [CrossRef]

Jurgovsky, Johannes, Michael Granitzer, Konstantin Ziegler, Sylvie Calabretto, Pierre-Edouard Portier, Liyun He-Guelton, and Olivier Caelen. 2018. Sequence classification for credit-card fraud detection. *Expert Systems with Applications* 100: 234–45. [CrossRef]

Kaggle. 2018. Credit Card Fraud Detection. Available online: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud (accessed on 27 July 2023).

Kim, Eunji, Jehyuk Lee, Hunsik Shin, Hoseong Yang, Sungzoon Cho, Seung kwan Nam, Youngmi Song, Jeong a Yoon, and Jong il Kim. 2019. Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning. *Expert Systems with Applications* 128: 214–24. [CrossRef]

Lan, Lan, Lei You, Zeyang Zhang, Zhiwei Fan, Weiling Zhao, Nianyin Zeng, Yidong Chen, and Xiaobo Zhou. 2020. Generative adversarial networks and its applications in biomedical informatics. *Frontiers in Public Health* 8: 164. [CrossRef]

Mariani, Giovanni, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. 2018. BAGAN: Data augmentation with balancing GAN. *arXiv* arXiv:1803.09655. [CrossRef]

Mirza, Shuja, Sonu Mittal, and Majid Zaman. 2018. Decision support predictive model for prognosis of diabetes using SMOTE and decision tree. *International Journal of Applied Engineering Research* 13: 9277–82.

Mo, Zan, Yanrong Gai, and Guanlong Fan. 2019. Credit card fraud classification based on GAN-AdaBoost-DT imbalanced classification algorithm. *Journal of Computer Applications* 39: 618–22.

Pradipta, Gede Angga, Retantyo Wardoyo, Aina Musdholifah, and I. Nyoman Hariyasa Sanjaya. 2021. Radius-SMOTE: A new oversampling technique of minority samples based on radius distance for learning from imbalanced data. *IEEE Access* 9: 74763–77. [CrossRef]

Scott, Mitchell, and Jo Plested. 2019. GAN-SMOTE: A generative adversarial network approach to synthetic minority oversampling. *Australian Journal of Intelligent Information Processing Systems* 15: 29–35.

Sharma, Anuraganand, Prabhat Kumar Singh, and Rohitash Chandra. 2022. SMOTified-GAN for class imbalanced pattern classification problems. *IEEE Access* 10: 30655–65. [CrossRef]

Tahir, Muhammad, Fazlullah Khan, Mohammad Khalid Imam Rahmani, and Vinh Truong Hoang. 2020. Discrimination of golgi proteins through efficient exploitation of hybrid feature spaces coupled with SMOTE and ensemble of support vector machine. *IEEE Access* 8: 206028–38. [CrossRef]

Xie, Yu, Guanjun Liu, Chungang Yan, Changjun Jiang, and MengChu Zhou. 2022. Time-aware attention-based gated network for credit card fraud detection by extracting transactional behaviors. *IEEE Transactions on Computational Social Systems* 10: 1004–16. [CrossRef]

Yang, Hao, and Yun Zhou. 2021. IDA-GAN: A novel imbalanced data augmentation GAN. Paper presented at the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, January 10–15; pp. 8299–305. [CrossRef]

Ye, Huanzhuo, Lin Xiang, and Yanping Gan. 2019. Detecting financial statement fraud using random forest with SMOTE. *IOP Conference Series: Materials Science and Engineering* 612: 052051. [CrossRef]

Zhang, Liyuan, Huamin Yang, and Zhengang Jiang. 2018. Imbalanced biomedical data classification using self-adaptive multilayer ELM combined with dynamic GAN. *Biomedical Engineering Online* 17: 181. [CrossRef] [PubMed]

Zhang, Zhaohui, Xinxin Zhou, Xiaobo Zhang, Lizhi Wang, and Pengwei Wang. 2018. A model based on convolutional neural network for online transaction fraud detection. *Security and Communication Networks* 2018: 5680264. [CrossRef]

Zhou, Tong, Jie Rong, Yang Liu, Weikang Gong, and Chunhua Li. 2022. An ensemble approach to predict binding hotspots in protein–RNA interactions based on SMOTE data balancing and random grouping feature selection strategies. *Bioinformatics* 38: 2452–58. [CrossRef] [PubMed]

Zhu, Xiaoqian, Xiang Ao, Zidi Qin, Yanpeng Chang, Yang Liu, Qing He, and Jianping Li. 2021. Intelligent financial fraud detection practices in post-pandemic era. *The Innovation* 2: 100176. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

International Journal of
Financial Studies

MDPI

Article

The Impact of Artificial Intelligence Disclosure on Financial Performance

Fadi Shehab Shiyyab, Abdallah Bader Alzoubi, Qais Mohammad Obidat and Hashem Alshurafat *

Department of Accounting, Business School, The Hashemite University, P.O. Box 330127, Zarqa 13133, Jordan; fadi_shiyyab@hu.edu.jo (F.S.S.); abdallahb@hu.edu.jo (A.B.A.); qais.mf.91@gmail.com (Q.M.O.)
* Correspondence: hashema@hu.edu.jo

Abstract: This study determines to what extent Jordanian banks refer to and use artificial intelligence (AI) technologies in their operation process and examines the impact of AI-related terms disclosure on financial performance. Content analysis is used to analyze the spread of AI and related information in the annual report textual data. Based on content analysis and regression analysis of data from 115 annual reports for 15 Jordanian banks listed in the Amman Stock Exchange for the period 2014 to 2021, the study reveals a consistent increase in the mention of AI-related terms disclosure since 2014. However, the level of AI-related disclosure remains weak for some banks, suggesting that Jordanian banks are still in the early stages of adopting and implementing AI technologies. The results indicate that AI-related keywords disclosure has an influence on banks' financial performance. AI has a positive effect on accounting performance in terms of ROA and ROE and a negative impact on total expenses, which supports the dominant view that AI improves revenue and reduces cost and is also consistent with past literature findings. This study contributes to the growing body of AI literature, specifically the literature on AI voluntary disclosure, in several aspects. First, it provides an objective measure of the uses of AI by formulating an AI disclosure index that captures the status of AI adoption in practice. Second, it provides insights into the relationship between AI disclosure and financial performance. Third, it supports policymakers', international authorities', and supervisory organizations' efforts to address AI disclosure issues and highlights the need for disclosure guidance requirements. Finally, it provides a contribution to banking sector practitioners who are transforming their operations using AI mechanisms and supports the need for more AI disclosure and informed decision making in a manner that aligns with the objectives of financial institutions.

Keywords: artificial intelligence; voluntary disclosure; content analysis; financial performance; banking sector; Jordanian banks

1. Introduction

Over the last decades, the study of artificial intelligence (AI) has involved the development of intelligent machines that can perform tasks requiring human intelligence. AI uses computer systems and algorithms to learn, reason, and make decisions based on data inputs. AI technologies mimic human cognitive abilities and can analyze data, automate tasks, and assist in various domains (Kok et al. 2002). AI refers to the scientific field and technology that involves the development of intelligent machines capable of imitating human behavior and intelligence (Ottosson and Westling 2020).

The significant advancements and practical applications of AI started to gain momentum in the 21st century with the advent of more powerful computing systems and the availability of large amounts of data (Ratia et al. 2018; Haenlein and Kaplan 2019). The specific implementation of AI will vary based on the industry, goals, and available resources. For example, to harness the potential of AI and gain a competitive edge in their respective industries, companies have developed in-house AI capabilities, partnered with AI solution providers, or utilized cloud-based AI platforms (Burström et al. 2021). The

banking sector is experiencing improved efficiency, accuracy, and personalized customer experiences as a result of AI implementation (Anastasi et al. 2021). AI offers numerous possibilities for banks to improve operations and drive innovation: data analysis, adaptive learning platforms, personalized marketing, automating repetitive tasks, chatbots, enabling natural language processing and voice recognition, and implementing risk-based predictive maintenance and fraud detection, among others.

AI is the process of human intelligence implemented by machines. AI promotes the sustainable and effective use of resources (Nikitas et al. 2020). Data-driven companies can enhance decisions and enable more precise predictions (Anastasi et al. 2021). Specifically, it is a more advanced digital transformation strategy that generates knowledge from existing large datasets (Lichtenthaler 2020). Thus, the implementation of AI processes will improve bank employees' productivity (Plastino and Purdy 2018). Previous studies indicate that banks have already recognized cost reduction and revenue generation through enhancing the quality of the operations process, for example, in terms of lending, security services, compliance improvements, fraud detection, and new types of services (Burgess 2017; Kaya 2019; Ryll et al. 2020). Moreover, these customized solutions and services provide customers with personalized investment strategies, wealth management techniques, and robo-advisors (Wheeler 2020). Currently, AI plays a vital role in autonomous decision-making processes, monitors assets and processes in real time, and enables value creation (Alcácer and Cruz-Machado 2019), and the benefits will increase going forward (Cockburn et al. 2018). In the rapidly evolving landscape of the banking sector, the integration of AI holds significant potential for enhancing decision-making processes and improving financial performance.

While AI can improve financial reporting, it can also lead to biases, lack of transparency, data privacy concerns, and compliance challenges. Organizations may face job displacement, training gaps, high implementation costs, interoperability challenges, and ethical concerns (Nguyen 2022). To mitigate these negative effects, organizations should prioritize responsible AI practices, invest in data quality and governance, and address potential biases in AI models. Staying informed about regulations and ethical considerations is also crucial (Nguyen and Dang 2023).

There are multiple motivations for conducting this study, including addressing stakeholders' concerns regarding the accountability and transparency of AI systems. Transparent disclosure can attract investors who value technologically informed decision making and potentially influence a company's valuation and shareholder composition. Aside from attracting investors, transparent AI disclosure can align with evolving regulatory frameworks. As regulatory bodies scrutinize the ethical and responsible integration of AI, companies that disclose their AI practices can demonstrate adherence to these guidelines, contributing to compliance and a robust corporate reputation (Meiryani et al. 2022).

These findings suggest that AI applications are favorable for the banking sector and beneficial for both shareholders and stakeholders, as well as for increased efficiency in the financial sector, which leads to economic benefits. However, quantifying the link between the use of AI and bank performance is warranted to explore the extent to which AI affects businesses, consumers, and the whole economy.

Despite the opportunities and benefits of the application of AI, AI disclosure is still voluntary. The decision of whether to disclose, to what extent, and the type of information is almost entirely left to the discretion of companies. To date, there is no commonly accepted practice for the level of AI disclosure. AI applications are relatively new. There are no known dedicated international reporting standards agreed upon in this area. The existing AI disclosure practices do not adequately capture the unique impacts of AI. The lack of a shared vision and reporting standards for AI leads to different disclosure practices depending on companies' perceptions (Sætra 2021).

While several studies have been conducted on the benefit of AI in the banking sector, there remains a research gap concerning the impact of AI disclosure on financial performance. Filling this gap is essential to shedding light on the potential benefits of AI

disclosure. To bridge this research gap, this study provides insights into the current level of AI-related term disclosure practices among Jordanian banks. We create an AI disclosure index by analyzing the spread of AI in the data of annual reports and investigating the impact of mentioning AI-related keywords on financial performance to explore the potential influence of disclosing AI-related terms on the financial performance of these banks.

Clear and ethical communication about AI initiatives can also enhance a company's image as a responsible innovator, fostering trust among consumers and partners (Hasan et al. 2023). On the other hand, a lack of transparency or negative perceptions could lead to skepticism and a tarnished reputation, potentially affecting financial performance. In addition, this study is motivated to provide valuable insights that contribute to the discourse on AI's role in shaping contemporary business success, offering a comprehensive perspective that encompasses ethical, regulatory, financial, and reputational considerations. Depending on the timing of the study, there could be a lack of empirical evidence regarding the direct relationship between AI disclosure and financial outcomes. Consequently, this paper aims to answer the following research question:

- Does artificial intelligence disclosure impact financial performance?

Based on a content analysis of 115 annual reports for 15 Jordanian-listed banks from the period 2014 to 2021, the results show an increase in AI-related keywords disclosure in the annual reports of Jordanian-listed banks from 2014 to 2021. The results also indicate that AI-related keywords disclosure has an influence on banks' financial performance. AI-related term disclosure has a positive effect on accounting performance in term of ROA and ROE and has a negative impact on the bank's total expenses, which supports the dominant view that AI improves revenue and reduces cost and is consistent with past literature findings.

This study contributes to the growing body of AI literature. First, it determines the spread of AI-related term disclosure in Jordanian banks by forming an initial AI-related term disclosure index. Second, it provides insights into the relationship between AI disclosure and financial performance. The findings of this study contribute to policymakers', international authorities', and supervisory organizations' efforts to address AI disclosure issues and highlight the need for disclosure guidance requirements. Moreover, the study provides a contribution to banking sector practitioners who are transforming their operations using AI mechanisms and supports the need for more AI disclosure and informed decision making in a manner that aligns with the objectives of financial institutions.

The remainder of this paper is organized as follows. The Section 2 reviews the relevant literature. The Section 3 illustrates the research methodology. The Section 4 presents and discusses the study findings. The Section 5 concludes the study.

2. Literature Review

2.1. AI Implications in the Banking Industry

The latest tech innovations that have revolutionized digital business are AI, machine learning, big data analytics, cloud computing, and social media. These innovations are used in the daily life of modern society. Technology changes physical, tangible items; contributes to operations enhancement; and promotes competence and ability for future business solutions (Tekic and Koroteev 2019). AI's abilities simplify its applications. Firstly, AI can forecast what is going on via the processing of audio, text, and computational linguistics in the surrounding environment. Secondly, AI relies on language and meaning through natural language processing (NLP) and assists humans in complying with machines through related AI algorithms. Finally, AI software systems can act alone without human contribution (Purdy and Daugherty 2016; Rao and Verweij 2017; Tákacs et al. 2018; Ottosson and Westling 2020). Moreover, unlike conventional machines, AI can even enhance itself all the time due to its self-learning ability based on past operation experiences (Öztemel and Gursev 2020). The new digital solutions are always changing the competitive strategies used in business and contributing to new approaches to value creation (Shang and Zhang 2022).

There is increasing feasibility of the utilization of chatbots by the banking industry. For example, a chatbot utilizes natural language technology to solve the user's problems (Suhel et al. 2020). The use of chatbots in the financial sector has changed the problem solving and answering of consumers' queries (Hwang and Kim 2021). Chatbots comprehend written and spoken text and can respond to vague questions and communicate with other portals or online data stores. Chatbot technologies handle massive numbers of calls from customers and enhance customers' satisfaction and confidence in banking services and their perceived usefulness (Sanny et al. 2020; Eren 2021; Nguyen et al. 2021). Moreover, chatbots handle more accounts than human advisors at lower operational costs and maximize profits (Patil and Kulkarni 2019). Furthermore, the 24/7 availability of online assessment chatbots has made operations more flexible, leading to a decrease in the use of physical bank branches (Wheeler 2020). Generally, banks utilize software such as UiPath, Automation Anywhere, Blue Prism, end-user devices, robots, software, and artificial intelligence agents to aid the process of repetitive banking operations (Vijai et al. 2020).

The introduction of AI enhances and simplifies the process of decision making while continuing to obey regulations. AI can reduce the number of false contracts, enhance the prediction of operational resources (Han et al. 2020), and accomplish obedience to regulatory requirements (Couchoro et al. 2021; Garcia-Bedoya et al. 2020; Kute et al. 2021). Numerous modern AI techniques have been utilized by banks to reduce fraudulent processes, such as data mining, fuzzy logic, machine learning, sequence alignment, and genetic programming (Raj and Portia 2011). Banks have improved their process speed, accuracy, and efficiency through the use of autonomous data management (Soni 2019). Predictive analytics can stop fraud incidents before they happen through several technologies such as Secure Socket Layer (SSL) for online transactions, encryption data storage, multi-level authorization, device fingerprinting, malware detection, token passwords, signing transactions, and endpoint protection (Kikan et al. 2019). Banks have utilized deep learning and artificial neural networks in personalized retail banking to assess their performance in direct marketing and determine which customers are likely to accept marketing proposals (Kim et al. 2015; Zakaryazad and Duman 2016). AI enhances the speed of the process, decreases cost-related issues, reduces operational risks, and improves know-your-customer processes through the use of chatbots and robot-advisors' services (Kaya 2019).

2.2. The Impact of AI on Financial Performance

The business community has witnessed an increased interest in the use of AI techniques, with financial services, manufacturing, information services, and banks as the most beneficial sectors (Bughin et al. 2017; Green et al. 2009). According to executives' surveys across various industries, investment in AI and focus on collaboration between employees and machine learning technology could increase banks' revenue by 34% (Shook and Knickrehm 2018).

Furthermore, according to the AI McKinsey Global Surveys series since 2004 and the current state of AI in 2023, AI adoption is increasing globally; the survey's results show an increase in companies embedding AI in at least one business function or one business unit.

Firm performance has received great attention from researchers in the accounting and finance area (Agarwal 2020). In general, researchers examine what might impact the financial performance of a firm in both positive and negative ways. For example, Almustafa et al. (2023) found that national governance quality plays an important role in mitigating the negative impact of the COVID-19 crisis on firm performance. Moreover, Nguyen (2022) found that new technologies, such as FinTech, might have a negative impact on financial performance. This finding was supported by Nguyen and Dang (2023), who found that FinTech development might negatively impact stock price and crash risk.

The Financial Stability Board (FSB) has analyzed the potential financial stability implications of the growing use of AI in financial services. FSB (2017) describes supply and demand factors that drive AI adoption in the financial sector; supply factors such as technological advances and the availability of data and infrastructure; and demand

factors such as profitability needs, competition with other firms, and the demands of financial regulation. According to the FSB, there are four sets of AI use cases: "(i) customer-focused (or 'front-office') uses, including credit scoring, insurance, and client-facing chatbots; (ii) operations-focused (or 'back-office') uses, including capital optimization, risk management model, market impact analysis, trading and portfolio management in financial markets; and (iv) uses of AI and machine learning by financial institutions for regulatory compliance ('RegTech') or by public authorities for supervision ('SupTech')" (FSB 2017).

Furthermore, the Organisation for Economic Co-operation and Development (OECD) has highlighted the potential impact of AI in specific financial market activities. According to the OECD (2021), the deployment of AI drives competitive advantages through two main avenues: first, by improving the firms' efficiency through cost reduction and productivity enhancement, therefore driving higher profitability (e.g., enhanced decision-making processes, automated execution, gains from improvements in risk management and regulatory compliance, back-office and other process optimization); and second, by enhancing the quality of financial services and products offered to consumers (e.g., new product offerings and high service customization). The use of AI generates new benefits, possibly leading to a substantial increase in labor productivity, operational workflow efficiency, and new revenue streams (PwC 2020). It will also strengthen risk management, improve customer experience, and enhance performance (Gokhale et al. 2019).

AI brings benefits to financial institutions in the form of stability, greater profitability, efficiencies in the provision of financial services, and systemic risk surveillance and regulation. From the firm's perspective, AI taking over repetitive bank tasks and autonomous AI solutions reduce the demand for less skilled labor and improve the efficiency of remaining staff (Kaya 2019). Thus, the implementation of speed-enhancing processes will improve bank employees' productivity (Plastino and Purdy 2018). Previous studies have indicated that banks have already recognized cost reduction and revenue generation through enhancing the quality of operations, for example, in terms of lending, security services, compliance improvements, fraud detection, and new types of services (Burgess 2017; Kaya 2019; Ryll et al. 2020). Moreover, these customized solutions and services provide customers with personalized investment strategies, wealth management techniques, and robo-advisors (Wheeler 2020).

These findings suggest that AI applications are favorable for the banking sector and beneficial for both shareholders and stakeholders, as well as for increased efficiency in the financial sector, which leads to economic benefits. The possible impact of AI in most practices has just started bringing benefits to financial institutions through various channels, improving efficiency and enhancing revenue sources by offering new services and products. Nevertheless, quantifying the link between the use of AI and bank performance is hard, not least due to issues around AI identification. In this context, a question arises about the extent to which AI affects businesses, consumers, and the whole economy.

2.3. AI Voluntary Disclosures and Financial Performance

The rapid advancements in AI have significantly impacted various sectors, and the banking industry is no exception. As banks increasingly integrate AI technologies into their operations, there is a crucial need to address the disclosure practices surrounding AI implementation (Sætra 2021). Financial services, among some other industries, with large volumes of customers and operations, depend on data processing in terms of text and speech; these services have most often adopted AI and disclose more information on natural language capabilities (AI McKinsey Global Surveys series).

Despite the opportunities and benefits of the application of AI, AI disclosure is still voluntary. The decision of whether to disclose, to what extent, and the type of information is almost entirely left to the discretion of companies. To date, there is no commonly accepted practice for the level of AI disclosure. AI applications are relatively new. There are no known dedicated international reporting standards agreed upon in this area. The existing AI disclosure practices do not adequately capture the unique impacts of AI. The lack

of a shared vision and reporting standards for AI leads to different disclosure practices depending on companies' perceptions (Sætra 2021).

In the age of Al, opacity is one of the greatest perils facing humans today. S. Lu (2021) argues that the current disclosure framework under corporate securities law has a limited impact on reducing opacity in Al systems. He calls for an effective disclosure framework for AI products and services in order to address AI opacity. It is crucial to ensure that AI algorithms are harmless by requiring more transparency to mitigate algorithmic opacity. Accordingly, such an advanced disclosure framework can help enhance transparency to reduce the risks posed to stakeholders, stabilize capital markets, and promote sustainability in the long run.

Recently, authorities' regulators and supervisors have been considering requirements around AI disclosure. For instance, the OECD addressed AI disclosure issues to promote innovative and trustworthy use of AI. Based on the OECD AI Principles, 'there should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them' (OECD 2019). Thus, financial consumers need to be informed about the use of AI techniques in the delivery of a product, as well as potential interaction with an AI system instead of a human being, to make conscious choices among competing products. Moreover, the OECD (2021) further highlights the need for disclosure requirements that help financial service providers to better assess whether prospective clients have a solid understanding of how the use of AI affects the delivery of the product. Similarly, the International Organization of Securities Commissions (IOSCO) also calls for including more clear information about the AI system's capabilities and limitations in such disclosures. IOSCO (2020) proposes guidance on "Appropriate transparency and disclosures to investors, regulators, and other relevant stakeholders". Disclosure requirements, therefore, need to consider any information around the use of AI techniques that may impact investors, regulators, and other relevant stakeholders, such as information on algorithmic trading models, data collection, and cross-border cooperation.

AI regulations are under development worldwide. There is a wave of new AI regulation that will soon have significant implications for AI systems. For example, the US "AI Disclosure Act of 2023: A Step Towards Algorithmic Transparency" represents an important step towards AI algorithmic transparency for those interacting with AI systems to be more informed about what they are interacting with, allowing them to make decisions accordingly. The scope of the AI Disclosure Act applies to any entity engaged in commerce, including banks (Section 5, the Federal Trade Commission). The EU AI Act (European Parliament 2023) also imposes transparency requirements for AI systems used in the EU. It is a significant piece of legislation that could have a major impact on the development and use of AI in the European Union. The EU AI Act specifies corresponding requirements for transparency, documentation, auditing, and obligations. AI systems should be developed to allow appropriate transparency and provision of information to users (Article 13). Such acts are still in the early stages of development, only having recently passed, but it is clear that they have the potential to shape the future of AI in the US and Europe. Becoming prepared early by establishing the appropriate disclosure procedures is the best way to ensure compliance with transparency requirements.

2.4. Development of Hypotheses

The prior literature discusses voluntary disclosure based on various theoretical perspectives. For example, agency theory focuses on the communication of specialized information by managers to enhance firm value and reduce costs of capital (Jensen and Meckling 1976). Voluntary disclosure allows managers to communicate effectively and enhance the firm's value since it encompasses various aspects, such as investment opportunities and financing policies. Agency theory scholars have called for expanding the categories of voluntary disclosure as a tool to observe the managers' actions and improve firm performance (Barako et al. 2006; Fang and Jin 2012; Cockburn et al. 2018; Haninun et al. 2018; Hassanein et al. 2019; Albitar et al. 2020). Similarly, the theory of capital needs points out that volun-

tary disclosure helps companies to increase additional funds at lower cost. Since companies need capital to continue their operations, disclosing more information voluntarily provides investors with additional information that enables them to make more valuable economic predictions about the firm (Bertomeu et al. 2011; Bini 2018; Cheynel 2013; Shehata 2014). From the perspective of signaling theory, companies would disclose more information than is mandatory in order to attract investors' attention, reduce information asymmetry between the company and its stakeholders, and signal their capabilities (Bertomeu et al. 2011; An et al. 2011). Managers voluntarily disclose good news to signal positive outcomes and may also disclose bad news to demonstrate efforts to address future losses. Managers of profitable firms prefer to voluntarily disclose significant information to signal profitability, boost investor confidence, and improve financial performance (Campbell et al. 2003; Albitar et al. 2020; Hassanein et al. 2019; Alkaraan et al. 2022). Legitimacy theory also offers a reasonable explanation for voluntary disclosure. Transparency, through voluntary disclosure, is one of the different strategies that companies adopt to legitimate their practices and activities to prove compliance with stakeholders' expectations (Magness 2006; Lightstone and Driscoll 2008; Bonsón et al. 2021). Organizations can gain social approval for their actions by engaging in higher levels of transparency (Bonsón and Bednárová 2013).

Regarding voluntary AI disclosure, companies want to inform stakeholders about increased operational efficiency to signal value creation, sustainability, and better competition in the market related to their AI products and processes. Thus, being more transparent about new AI technologies would attract potential investors and increase the trustworthiness of other stakeholders. Information about applied AI tools is communicated to different stakeholders to gain their trust and acceptance and obtain legitimation (Osburg 2017; Thorun 2018; Bonsón et al. 2021, 2023). Similar to other voluntary disclosure challenges, the applications of AI models raises important disclosure issues in the banking industry due to the nature of the business and its operational environment. The growing development and utilization of AI can influence voluntary disclosure in the banking industry (Saenz et al. 2020). The competitive landscape in the banking industry impacts the decisions to disclose information about firms' use of AI. Firms may elect to disclose more voluntary information if they believe that will gain a competitive advantage. For example, banks like to show their expertise in the field (Yu et al. 2017; Krakowski et al. 2022). Therefore, banks that operate in highly competitive markets may choose to disclose more information about their AI systems to be competitive with their competitors. According to the AI McKinsey Global Surveys series, companies are most likely to adopt AI in functions that provide core value in their industry, and financial services are more likely than other industries to have adopted AI in service operations, marketing, sales, and risk functions.

AI is found to have a positive impact on a firm's nonfinancial indicators, such as the accounting information system efficiency and the success of the workflow within a firm (Hashem and Alqatamin 2021). Wamba-Taguimdje et al. (2020) stated that AI benefits organizations, and more specifically, their ability to improve performance at both the organizational (financial, marketing, and administrative) and process levels. Fethi and Pasiouras (2010) provided a comprehensive review of 179 academic papers that examined the impact of AI on different financial indicators of banks. They concluded that one of the financial indicators that is affected positively by AI is the banks' financial performance. Therefore, the following first hypothesis is suggested:

H1. *AI-related terms disclosure will be positively related to the banks' financial performance.*

In the same vein, banks that prefer clarity, moral behavior, transparency, and social responsibility are more likely to disclose information about their AI systems if they believe that disclosure enhances confidence and credibility among customers, regulators, and society at large (ElKelish and Hassan 2014). Disclosing information about AI practices, their algorithm models, data sources, and decision-making operations enhances justice and trust among customers, regulators, and stakeholders. Moreover, it keeps banks as far as possible from any reputational dangers that may emerge from AI utilization decisions. For

example, companies that have confidence in their use of AI are more likely to disclose more related information. Disclosing information about AI use creates a reputation of innovation and excellence that engages more clients who value advanced AI systems. AI transparency enhances banks' reputation, customer loyalty, and profitability (Felzmann et al. 2020). In addition, the banking industry is highly sensitive to stricter regulations in contrast to other industries. The banks may disclose more information about their AI systems to show their compliance with moral standards and gain a competitive edge in markets that value responsibility (Krakowski et al. 2022). Overall, the tendency of AI voluntary disclosure is to highlight favorable information to obtain positive effects in terms of economic impact and to appeal to investors, as this disclosure can help investors gain a clearer picture in terms of the company's investments in AI technologies (Bonsón et al. 2021).

Based on various theoretical perspectives, expectations, and the supporting literature, it is assumed that companies that apply AI in their processes and recognize a financial benefit already, in terms of increased revenue and decreased costs, are motivated to disclose positive information about the development, application, and use of AI. It is also anticipated that companies that voluntarily disclose AI information more often are more likely to have implemented AI as part of their business strategy and recognized financial benefits. Hence, we expect the frequency of disclosure of AI-related terms in banks' annual reports to impact financial performance in terms of increased revenue and decreased costs compared to other banks.

The advancement of new technological solutions such as AI has shown its merit through the impact on organizations' different financial indicators, and the total cost is one financial indicator that is expected to be reduced by the adoption of AI techniques (Nguyen and Dang 2023). Fethi and Pasiouras (2010) indicated that the bank's total cost would be minimized by using AI techniques. This evidence is supported by Doumpos et al. (2023), who reviewed AI techniques in the banking literature and concluded that AI techniques are expected to enhance financial indicators within the banking industry. Therefore, the following second hypothesis is suggested:

H2. *AI-related terms disclosure will be negatively related to the banks' total expenses.*

3. Research Design

3.1. AI Disclosure

Following the purpose of this study, we use the content analysis method for measuring levels of AI disclosure and produce a preliminary list of AI keywords. Previous studies have applied similar processes (e.g., Hassanein et al. 2019; Elmarzouky et al. 2021; Karim et al. 2021; Alkaraan et al. 2022). In particular, the selection of AI-disclosure-related items is carried out in three stages.

Firstly, we create an AI disclosure index through a comprehensive review of AI components that have been mostly mentioned in the finance sector by related professional organizations such as the FSB (2017), OECD (2019), and IOSCO (2020). For example, the AI Index Report 2019 highlight the terms most often mentioned in the finance sector, including "AI", "Big Data", "Cloud", and "Machine Learning" (Perrault et al. 2019), based on the previously related literature that utilizes various self-constructed disclosure proxies' measures (Finkenwirth 2021; Zetzsche et al. 2020; Hussainey et al. 2022). In the context of AI, Omar et al. (2017) applied content analysis to the annual reports of Malaysian publicly listed companies, searching for the words "Artificial Intelligence", "Machine Learning", and "Big Data". Cam et al. (2019) conducted a content analysis and classified the AI applications in USA banks; "Robotic Process Automation", "virtual agents", "natural language text understanding", "machine learning", and "computer vision" are mentioned frequently. Bonsón et al. (2023) conducted content analysis of European listed companies, and they identified search keywords items including "artificial intelligence", "machine learning", "automat", and "algorithm".

According to McWaters (2018), focusing on AI alone is not sufficient to understand the myriad ways in which it could be used within financial institutions. Advances in any

technology will increase the capabilities of all other technologies that interact with it. Thus, AI must be understood within the context of all other technologies. We combined a set of the most often mentioned AI-related terms in the finance sector to identify the preliminary AI disclosure index. Appendix A Table A1 Column A presents the keywords for AI-related disclosure terms.

Second, the frequency of AI-related terms for each bank's annual report is measured through the computerized content analysis software "Maxqda". The keywords and the context of those keywords are also analysed, including the sentence or sentences before and after the search term, to provide insight into the AI business strategy.

Third, the AI-related terms disclosure keywords are classified into three categories. The first group combines the keywords related to digital awareness, transformation, and capabilities. The second group is related to AI applications, products, services, and processes. The last group is related to AI challenges and threats in terms of information and cyber security (Appendix A Table A1 Column B presents the classes keywords for AI-related disclosure terms for each category). Thereafter, the resulting data for the content analysis are incorporated into multiple regressions.

3.2. Regression Model and Variable Definitions

The following regression model is used to measure the impact of mentioning AI-related keywords in the annual reports on financial performance.

$$\begin{aligned} \text{Performance t} = {} & ß_0 + ß_1\ \text{AIFREC}_{t-1} + ß_2\ \text{BSIZE}_{t-1} + ß_3\ \text{DEBT}_{t-1} + ß_4\ \text{BDSIZ}_{t-1} \\ & + ß_5\ \text{INDPB}_{t-1} + ß_6\ \text{FORSH}_{t-1} + ß_7\ \text{LASHR}_{t-1} + ß_8\ \text{BRNCH}_{t-1} \\ & + ß_9\ \text{BKAGE}_{t-1} + ß_{10}\ \text{YEAR} + e \end{aligned} \quad (1)$$

where performance is the bank's financial performance, assessed by several alternative estimations. We consider accounting measures of performance, including return on equity (ROE), return on assets (ROA), and net interest income (NII), in our basic analysis. We also consider market performance measures, such as price earnings ratio (P/E). ROA and ROE are widely used in the literature as the more effectively a company's management produces revenue from its assets and from shareholders' capital, the higher they are; thus, the higher the ROA and ROE, the higher the firm performance (Almustafa et al. 2023; Hasan et al. 2023). Alternatively, we also consider total expenses (TEXP), assuming that AI implementation reduces cost, which leads to higher financial performance. NII reflects the main banking operation's income from lending and borrowing, which is assumed to be affected directly by increased revenue and decreased cost of these main services. These ratios provide a picture of the firm's financial development; past studies have used them for evaluating financial performance (Hagel et al. 2013; Heikal et al. 2014; Fatihudin 2018).

AIFREC ('AI frequency') is the AI-related terms disclosure frequency measured as the number of mentioned AI-related terms in each annual report (Alkaraan et al. 2022; Finkenwirth 2021). The AI practices and AI disclosure choices at each bank are sensitive to many internal and external factors. The banks' specific characteristics create incentives for bank managers for AI implementation and AI disclosure practices according to these features. The literature provides abundant research on the association between banks' characteristics and bank performance. In our empirical models, we include three groups of control variables that reflect banks' governance features, ownership structure, and cost of capital attributes, and economic-specific characteristics, many of which are commonly known from the previous literature to be important factors affecting the banks' performance measures.

The first group of control variables reflects the internal corporate governance features at the bank level. We control for board size and board independence (Jensen 1993; Yermack 1996). Independent directors are deemed more effective monitors due to greater reputational costs (Fama and Jensen 1983; Coles et al. 2008). The board of directors is considered a core internal corporate governance variable in the literature (see S. Agarwal (2020) for a review; Dang and Nguyen 2021; Nguyen 2022). Hence, board size (BODSIZE) is the number of board directors, and independent board (INDD) is the proportion of

independent directors from the board. By including this control, we cater for the case that a bank's performance and AI disclosure are products of the board on which they serve.

In the second group, we include controls for bank ownership structure and costs of capital attribute variables that may affect performance. Theoretically, agency costs of capital arise from the conflict between shareholders and debt holders of a public company. The debtholders may also place limits on the use of their capital if they believe that management will take actions that favor shareholders instead of debtholders. Thus, we control the bank ownership structure, including the large shareholders' ratio (LSHAR), which is measured as the percentage of accumulated large shareholders who own 5% or more of the bank's stock. Foreign ownership ratio (FORSH) is used to control for investors' interest (Al-Gamrh et al. 2020; Mallinguh et al. 2020; Nguyen 2022). We also control for the debtholders' structure. The literature has found that debt ratio affects financial performance (Almustafa et al. 2023; Gander 2012; Shiyyab et al. 2014). The debt ratio (DEBT) is measured as the ratio of long-term debt to total debt. Large shareholders are predicted to be associated with AI disclosure and performance.

Third, we control for economic characteristics at the bank level, consistent with what has been documented in the bank performance literature. First, it has been documented that bank size and operation complexity have an impact on management activities and, primarily, bank performance (Almustafa et al. 2023; Nguyen 2022). Therefore, we control bank size as the natural logarithm of total assets (BSIZE). In addition, we control for the number of branches and bank age. Despite the widespread use of online banking services, the banks' physical presence is still important due to the opportunities for face-to-face contact with the customer (Almustafa et al. 2023). The literature provides evidence that branch networks still enable the bank to gain a competitive advantage by creating extensive-term personalized links with their clients, leading to increased profits and reduced loan losses (Berger and Black 2011; Hirtle 2007; Harimaya and Kondo 2012; Kondo 2018; Monferrer Tirado et al. 2019). Therefore, we control for the number of branches with BRNCH as a continuous variable reflecting the number of branches for each bank every year, and we also include a control for bank age (BKAGE) as the total number of years a corporation has been in operation (Almustafa et al. 2023; Nguyen 2022; Coad et al. 2018).

Finally, we also account for the impact of the COVID-19 crisis on bank performance; we use a dummy variable as the independent variable, which is 1 for the COVID-19 period or 0 otherwise. We define the COVID-19 crisis period as 2020 (Shen et al. 2020; Almustafa et al. 2023; Hasan et al. 2023). Finally, we also control for years' effects. We created an indicator variable representing the year-dummies (YEAR_DUM) to control for year-specific effects. All variable definitions are presented in Table 1.

Table 1. The variables' names and definitions.

	Variables Definitions
AIFREC	The number of mentioned AI-related terms in each annual report
ROA	Return on assets
ROE	Return on equity
NII	Net interest income
P/E	Price-to-earnings ratio
TEXP	The natural logarithm of total expenses
BKSIZ	The natural logarithm of assets
DEBT	The long-term debt to total debt
BDSIZ	The number of board directors
INDPB	The proportion of independent directors on the board
FORSH	The ratio of foreign shares to total shares
LASHR	The percentage of large shareholders who own more than 5% of total shares
BRNCH	The number of branches for each bank every year
BKAGE	The total number of years from the date of establishment of the bank
COVID-19	A dummy variable, equal to 1 for the COVID-19 period (the year 2020) or 0 otherwise

3.3. Sample and Data

Financial analysts use both financial and non-financial information to gain a reliable picture of companies' performance, and the annual report is one of the main sources for the decision-making process of investors in the financial market (Araújo Júnior et al. 2014; Zhou et al. 2017). Therefore, this research is based on the analysis of 130 annual reports for all 15 Jordanian-listed banks from the period 2014–2022. Most of the annual reports are available as a PDF version for all Jordanian banks on their websites. The relevant words are searched, being those previously mentioned in the context analysis. The newly created keyword dataset based on the annual reports provides insights into the development of AI-related mentions.

4. Finding and Discussion

4.1. Descriptive Statistics

Table 2 panel A provides an overview of the AI-related terms disclosure for each bank. The total AI-related keywords disclosure is 2658, while he range is wide between banks. The highest AI-related keywords disclosure 18% of the dataset belongs to the Jordan Ahli Bank, 12% to Safwa Islamic Bank, and 11% to Bank El Etihad. These banks mention AI-related terms with highly frequency, while Societe General, International Islamic Arab Bank, and Invest Bank are banks that mention AI-related terms with low frequency. The remaining banks' samples range from 4% to 9% of the dataset. Table 2 panel B presents the frequency of AI-related keywords mentioned in an annual report by year. The AI-related keywords disclosure varied across years and increased dramatically from 2014 to 2022 (733 related words in 2022, 73 in 2014). However, approximately 60% of AI-related keywords were disclosed in the last three years of 2019-2022. This reflects the recent revolution of AI in the banking industry, especially during COVID-19 time. The increase in AI-related disclosure in annual reports suggests that banks have made more investments and rely more on AI implementation.

Table 2. Summary statistics of AI disclosure by bank and by year.

Panel A. AI Disclosure Frequency by Bank			Panel B. AI Disclosure Frequency by Year		
Bank Name	**AI Freq**	**Per%**	**Year**	**AI Freq**	**Per%**
Jordan Ahli Bank	471	18%	2014	73	3%
Safwa Islamic Bank	311	12%	2015	75	3%
Bank El Etihad	283	11%	2016	122	5%
Bank of Jordan	228	9%	2017	220	8%
The housing Bank for Trade	215	8%	2018	289	11%
Arab Bank	211	8%	2019	297	11%
Jordan Commercial Bank	172	6%	2020	441	17%
Cairo Amman Bank	165	6%	2021	515	19%
Jordan Kuwait Bank	137	5%	2022	733	23%
Jordan Islamic Bank	118	4%			
Arab Banking Corporation	113	4%			
Arab Jordan Investment Bank	103	4%			
Societe General Bank	54	2%			
International Islamic Arab Bank	48	2%			
Invest Bank	29	1%			
Total	2658	100%		2658	100%

The AI-related disclosure keywords are classified into three categories. The first group combines the keywords related to digital awareness, transformation, and capabilities. The second group is related to AI applications, products, services, and processes. The last group relates to AI challenges and threats in terms of information and cyber security. Table 3 presents the AI-related disclosure terms classified into three categories.

Table 3. The AI-related disclosure terms are classified into three categories.

AI-Related Terms/Words	Frequency	Percentage
AI digital awareness, transformation, and capabilities	640	24.09%
AI application, product, service, and process	886	33.33%
Information and cyber security	1132	42.58%
Total—AI-related terms/words	1925	100%

The first group combining the keywords related to digital awareness, transformation, and capabilities represents 24% of the total, which highlights the banks' keenness and commitment to harnessing the potential of artificial intelligence in their operations and services. The most frequently recurring keywords are digital transformation (143), fintech (130), and financial technology (85). The annual reports of banks have mentioned artificial intelligence 13 times, reflecting a significant interest and focus on this transformative technology. This indicates a potential awareness within banks' management regarding the pivotal role these terms play in attracting attention to the banks' pioneering role and achieving a competitive advantage.

The second group related to AI applications, products, services, and processes represents 33% of the total. This reflects the level of dedication towards essential services in achieving digital transformation and utilizing artificial intelligence technologies. This signifies the unwavering commitment to harnessing the power of cutting-edge technologies and embracing a future where innovation and digital advancements drive progress. The keywords "robotic automation", "digital banking", "mobile banking", "online banking", and "digital services" are of the most importance in the banking industry. This signifies the transformation and advancements that are shaping the future of banking. Robotic automation streamlines operations, while digital banking, mobile banking, and online banking offer convenient access to services. Digital services encompass innovative solutions that enhance the overall banking experience. Together, these keywords represent the vital role technology plays in revolutionizing the industry.

The last group is related to AI challenges and threats in terms of information and cyber security. The extent of interest in this area is evident in the remarkable percentage (44%) highlighting collective challenges and threats. The most frequently mentioned keywords in this group are information security, cyber security, electronic security risks, IT security, electronic banking services, and electronic security policies. This finding demonstrates a strong emphasis on ensuring robust security measures in the digital landscape. These keywords highlight the importance of protecting sensitive data, combating cyber threats, and maintaining secure electronic banking services. They signify a commitment to safeguarding digital systems and providing secure experiences for individuals and organizations. This is consistent with previous results and industry surveys. For example, the global joint survey conducted by the World Economic Forum and the Cambridge Centre for Alternative Finance indicates that a majority of financial services companies (56%) have implemented AI technology in terms of risk management domains.

Table 4 provides descriptive statistics for the total sample. The mean of the AI-related terms frequency variable is 20.98, with a minimum of 2 and a maximum of 121. This reflects that some banks need more disclosure information with regard to AI-related terms. The bank characteristics also vary according to each bank; for example, BKSIZ varies between 22 and 8441 million, BRNCH ranges between 12 and 211 branches, and the bank age ranges between 5 and 88 years of operation. These differences are expected to explain bank performance as well as AI disclosure.

Table 4. Descriptive statistics.

Variable	Obs No.	Mean	Median	Sta Dev	Min	Max
AIFREC	127.00	20.98	14.00	20.63	2.00	121.00
ROE	127.00	7.84	7.92	4.36	−1.42	16.50
ROA	127.00	0.94	0.99	0.52	−0.43	2.05
P/E	127.00	12.87	10.93	9.33	0.00	62.10
NII (in millions)	127.00	174.48	111.89	249.17	9.11	913.23
TEXP (in millions)	127.00	109.71	63.20	165.02	11.78	860.18
BKSIZ (in millions)	127.00	2586	2200	1950	22.00	8441
DEBT	127.00	87.58	87.64	2.92	81.71	93.04
FORSH	127.00	0.07	0.05	0.09	0.00	0.52
LARSHR	127.00	0.67	0.66	0.21	0.30	0.97
BODSIZ	127.00	11.46	12.00	1.88	5.00	13.00
INDPB	127.00	0.41	0.42	0.11	0.18	0.64
BRNCH	127.00	64.78	48.00	49.61	12.00	211.00
BKAGE	127.00	40.69	39.00	18.81	5.00	88.00

AIFREC is the frequency of mentioned AI-related terms in each annual report. ROA is return on assets. ROE is return on equity. NII is the natural logarithm of net interest income. P/E is the price-to-earnings ratio. TEXP is the natural logarithm of total expenses. BDSIZ is the number of board directors. INDPB is the proportion of independent directors from the board. LASHR is the accumulated percentage of large shareholders who own 5% or more of the bank's stock. FORSH is the ratio of foreign shares to total shares. DEBT is the ratio of long-term debt to total assets. BKSIZ is the natural logarithm of assets. BRNCH is a number of branches for each bank every year. BKAGE is the number of years a corporation has been in operation.

4.2. Correlations

Table 5 provides the results of the correlations among variables, with some coefficients warranting particular attention. Overall, the correlations are relatively small, and the low inter-correlations among all independent variables indicate that multicollinearity does not appear to be a problem in the regression model. However, it provides valuable information regarding the associations between AI disclosure and firm variables. For example, we observe several associations between AI disclosure and bank characteristics that identify which bank features are more conducive to AI disclosure practices. AI disclosure practices are positively associated with corporate governance features in terms of BDSIZ and INDPB (at 1% and 10%, respectively). Similarly, as expected, AI disclosure is positively associated with BKAGE %, BKSIZ 5%, and BRNCH 10%. FORSH is positively associated with AI disclosure, whereas LASHR is negatively correlated with AI disclosure, which indicates that shareholders are either conservative with regard to AI implementation or they have symmetric information, which implies weak disclosure in the annual reports. LASHR is positively associated with BKSIZ, BDSIZ, and INDPB, while it is negatively related to FORSH. DEBTH is negatively correlated with FORSH, BRNCH, and BKAGE. BDSIZ is associated with INDPB and BRNCH. As expected, the economic characteristics are correlated to each other; BKSIZ is linked to BRNCH and BKAGE, all at a 1% significance level, and to the BDSIZ at a 5% level of significance.

Table 5. Correlations.

	AIFRCD	BKSIZ	DEBTH	LASHR	FORSH	BDSIZ	INDPB	BRNCH
AIFREC	1.000							
BKSIZ	0.220 **	1.000						
DEBTH	0.107	−0.032	1.000					
LASHR	−0.327 ***	0.178 **	0.048	1.000				
FORSH	0.247 ***	0.266 ***	−0.205 **	−0.188 **	1.000			
BDSIZ	0.234 ***	0.173 *	−0.283	0.306 ***	−0.089	1.000		
INDPB	0.149 *	−0.041	0.175	0.199 **	0.073	0.170 *	1.000	
BRNCH	0.153 *	0.915 ***	−0.155 *	0.132	0.197 **	0.218 **	−0.192 **	1.000
BKAGE	0.275 ***	0.684 ***	−0.294 ***	−0.087	0.151 *	0.322 ***	−0.038	0.763 ***

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.3. Regression Model

Multiple regression analysis using ordinary least squares (OLS) has been used. Table 6 reports the results of OLS regressions of AI-related term disclosure on a set of performance indicator variables. Models 1, 2, 3, and 4 present the results for ROA, ROE, NII, and P/E ratio, respectively. Model 5 presents the result for total expenses. The models have a predictive capacity for the dependent variables, in terms of R2, which ranges between 0.27 and 0.64, and F-values are significant at the 1% level.

Table 6. The OLS regressions of AI disclosure on financial performance.

	(1)	(2)	(3)	(4)	(5)
	ROA	**ROE**	**P/E**	**NII**	**TEXP**
AIFREC	0.00427 **	0.0314 **	0.0567	0.0172 *	−0.00355 **
	(2.86)	(2.54)	(1.25)	(1.97)	(2.31)
LASHR	−1.338 ***	−11.48 ***	3.059	−2.376 ***	−0.0298
	(−5.49)	(−5.75)	(0.54)	(−4.44)	(−0.16)
DEBTH	−0.0382 **	−0.284 **	−0.204	−0.0850 **	−0.0353 ***
	(−2.56)	(−2.32)	(−0.59)	(−2.59)	(−3.03)
FORSH	1.312 ***	11.44 ***	−18.55	−0.118	1.028 ***
	(3.03)	(3.22)	(−1.86)	(−0.12)	(3.05)
BDSIZ	0.0226	0.0619	0.266	0.352 ***	−0.00208
	(0.89)	(0.30)	(0.45)	(6.30)	(−0.10)
INDPB	−1.401 ***	−12.80 ***	7.738	−0.184	0.0754
	(−3.81)	(−4.25)	(0.91)	(−0.23)	(0.26)
BKSIZ	−0.0168	−0.855	−0.194 *	−0.498 **	0.314 ***
	(−0.15)	(−0.95)	(−1.58)	(−2.05)	(3.66)
BRNCH	0.00591 **	0.0763 ***	0.0458	0.0493 ***	0.00913 ***
	(2.57)	(4.05)	(0.86)	(9.76)	(5.10)
BKAGE	−0.0158 ***	−0.180 ***	0.0319	−0.0865 ***	0.00697 ***
	(−3.78)	(−5.26)	(0.33)	(−9.44)	(2.45)
Year Dummy	Yes	Yes	Yes	Yes	Yes
Bank Dummy	Yes	Yes	Yes	Yes	Yes
_cons	6.123 **	15.97 ***	22.56 *	19.59 ***	13.21 ***
	(2.55)	(3.81)	(1.91)	(3.72)	(7.08)
N	127	127	127	127	127
R^2	0.466	0.495	0.278	0.632	0.641
adj. R^2	0.425	0.456	0.212	0.579	0.594

t-statistics in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. AIFREC is the frequency of mentioned AI-related terms in each annual report. ROA is return on assets. ROE is return on equity. NII is the natural logarithm of net interest income. P/E is the price-to-earnings ratio. TEXP is the natural logarithm of total expenses. BDSIZ is the number of board directors. INDPB is the proportion of independent directors from the board. LASHR is the accumulated percentage of large shareholders who own 5% or more of the bank's stock. FORSH is the ratio of foreign shares to total shares. DEBT is the ratio of long-term debt to total assets. BKSIZ is the natural logarithm of assets. BRNCH is the number of branches for each bank every year. BKAGE is the number of years a corporation has been in operation.

Models 1 and 2 show that performance indicator variables such as ROA and ROE provide an explanation for AI-related term disclosure. In particular, the impact of AI-related term disclosure is positive and statistically significant, with ROA and ROE at a 5% level of significance. This indicates that additional AI-related term disclosure is valued and related to better performance. Moreover, AI-related term disclosure provides an additional signal for specialist technology knowledge and expertise in modern operational settings that allow for additional performance. These results signify that AI frequency has an influence on the bank's profitability and shareholders' equity. In the same vein, the impact of the frequency of AI-related term disclosure on total expenses (TEXP) is negative and statistically significant at a 5% level of significance, which supports the dominant view that AI reduces cost and is consistent with past findings. For example, banks are already

strengthening customer relationships and lowering costs by using artificial intelligence (McKinsey & Company 2021).

The bank-specific features also provide additional insight into banking performance and can explain the variation between banks' performances. For example, FORSH is positively and statistically significantly correlated with bank performance in terms of ROA and ROE at a 1% level of significance. This result is consistent with previous studies that found foreign ownership positively affects firms' financial performance (Al-Gamrh et al. 2020; Mallinguh et al. 2020) and financial stability in emerging markets (Nguyen 2022). BRNCH is positively associated with bank performance in terms of ROA, ROE, and NII. BDSIZ is positively and statistically significantly correlated with NII at a 1% level of significance, while it did not have any significant influences on other measures of performance.

LASHR is negatively and statistically significantly correlated with bank performance in terms of ROA, ROE, and NII at a 1% level of significance. This finding is consistent with previous studies that found large and concentrated shareholdings have a negative impact on performance (Al-Malkawi et al. 2014; Abdallah and Ismail 2017). DEBTH is negatively and statistically significantly correlated with bank performance in terms of ROA, ROE, and NII at a 1% level of significance. This is consistent with Almustafa et al. (2023)'s finding that DEBTH is negatively and statistically significantly correlated with ROE and ROA.

Unexpectedly, BKAGE is negatively and statistically significantly correlated with the bank performance measures ROA, ROE, and NII. Loderer et al. (2017) argue that firm age increases organizational rigidity, monitoring, and corporate control, thus leading to declining growth opportunities; they found evidence that companies invested less as they grew older. Similarly, INDPB is negatively and statistically significantly correlated with the bank performance measures ROA and ROE. BKSIZ is also negatively and statistically significantly correlated with bank performance measures in terms of PE and NII. This indicates that the too-big-to-fail problem may exist. The results are consistent with previous studies that found large firms reduce financial stability and diversification in emerging markets (Nguyen 2022) and are associated with more risk due to the "too big to fail" problem (Zardkoohi et al. 2018; Almustafa et al. 2023; Nguyen 2022). Model 5 in Table 6 shows that TEXP is positively and statistically significantly correlated, all at a 1% level, with each of FORSH, BKSIZ, BRNCH, and BKAGE. However, TEXP is negatively and statistically significantly correlated with DEBTH, at a 1% significance level.

Finally, we investigate the impact of the COVID-19 crisis on firm performance. Firstly, we use a dummy variable, which is 1 for the COVID-19 period or 0 otherwise. We further split the sample into two parts: before and during COVID-19. Overall, we do not observe any impact on the relationship between AI disclosure and bank performance. Hasan et al. (2023)'s results indicate that COVID-19 does not necessarily significantly impact business performance outcomes. They argue that modern technology, such as artificial intelligence, significantly mitigates the negative impacts. Similarly, Almustafa et al. (2023) argue that the national governance system has significantly reduced the impact of the COVID-19 crisis on firms' operations, and government support reduces the effects of economic shock, especially in banking sectors. Meanwhile, some variables' coefficients have been changed as expected due to the COVID-19 crisis impact, as seen in Table A5 in Appendix C.

Although some evidence is provided that AI adoption has a positive effect on accounting performance, other performance indicator variables lack significance. These results are consistent with Nguyen (2022)'s finding that FinTech development negatively affects financial stability in an emerging market. A possible interpretation is that annual reports do not provide enough insights into AI implementation. It might be that banks implementing AI have not mentioned AI-related terms in their annual reports due to a lack of AI disclosure requirements. Therefore, banks that mentioned fewer AI-related terms have not adopted AI yet or adopted AI in limited business units. Moreover, the disclosure of AI-related terms is not enough to achieve financial impact, as the benefit of adopting AI cannot be expected automatically based solely on the AI disclosure mentions.

The weak relation may be due to the contextual settings in well-established developed customer relations countries, such that AI disclosure may not fully affect performance. If the banking operation's performance is highly developed, stable, and automated, AI disclosure may not require changes in some banking practices. These AI disclosure benefits could have had little impact on the bank's performance.

In line with these findings, the result shows that Jordanian-listed banks increased AI-related keywords disclosure in their annual reports. Banks developing AI applications often mention use cases and point out the benefits of AI for improving their operation. This indicates the importance of integrating AI technologies into business models, leading to lower cost and higher performance, which is consistent with previous AI results and industry surveys. Companies have already recognized the contribution of AI adoption for better overall performance, increased revenue, and decreased cost (AI McKinsey Global Surveys series; the global joint survey conducted by the World Economic Forum and the Cambridge Centre for Alternative Finance). Overall, the results align with the tendency of various theoretical perspectives of AI voluntary disclosure that companies are motivated to highlight favorable information to obtain positive economic impact and to appeal to investors.

5. Conclusions

This study examined whether AI-related references in annual reports could be used as an explanatory variable for financial performance. We analyzed 115 annual reports for 15 Jordanian-listed banks from the period 2014–2021. The analysis of annual reports shows an increase in the frequency of AI-related terms disclosures since 2014. This development indicates that Jordanian banks have become more aware of AI adoption, implications, and benefits. At the same time, there is a weak level of AI-related disclosure in some Jordanian banks, which indicates that they are still at an early AI implementation stage, at least on the level of AI disclosure. As the trend of AI adoption is still developing, more efforts are needed for improvement in the context of voluntary AI disclosures.

Based on the results, the presence of AI-related keywords in a bank's disclosures positively impacts its profitability and efficiency, as indicated by improved ROA and ROE. It also leads to a decrease in total expenses, suggesting that AI is streamlining operational processes and reducing costs. These findings demonstrate AI's potential to drive revenue growth and enhance efficiency in the banking sector.

Based on AI-related mentions in the annual reports of Jordanian banks, this study shows a positive impact of AI-related term disclosure on accounting performance, and financial benefits have been realized. To the best of the researchers' knowledge, this study is the first in Jordan that links AI-related terms disclosure in annual reports to financial performance. This study contributes to the existing literature by providing new evidence of AI voluntary disclosures, specifically offering insights into how Jordanian banks disclose information related to AI in their annual reports.

This study provides bank executives, annual report users, regulators, and policymakers with a view of AI disclosure's impact on financial performance and the competitive advantage of AI disclosure in financial services. The study's findings are relevant to annual report providers in that disclosures related to AI are increasing in the banking industry and are of interest to users. In particular, AI disclosure might be useful to investors and financial analysts because it helps them to gain a clearer picture in terms of the company's investments in AI technologies and the sustainability of their investments. The study also provides regulators with recent evidence on voluntary disclosures in general and disclosures on AI that can help regulators assess current trends in AI voluntary disclosures, understand the challenges and opportunities of AI, and predict future directions in the adoption and management of AI. Regarding implications for policymakers, we highlight the importance of establishing a unified AI disclosure framework, making annual reports more transparent and easier to understand for investors and other stakeholders. As a result,

we support the new AI regulation development worldwide that enhances the quality and clarity of the AI information presented in annual reports.

According to our results, we provide a topic for future research. The AI adoption decisions and AI-related terms disclosures may be driven by company culture, corporate governance, top management leadership, and ownership structure. Future research may consider more firms' characteristics and control for other factors that drive the success of AI implications, adoption, and disclosures. In addition, the benefit of AI implementation could be reviewed on the business unit level or process level rather than the level of the banks' overall performance. Future research could consider the business unit performance separately.

The results of this study have substantial practical implications for banks and their financial performance. This study shows that banks can improve their financial performance by voluntarily disclosing their AI initiatives. This increases stakeholder trust and attracts AI-informed investors. It also helps mitigate risks associated with AI and gives banks a competitive advantage through differentiation. Being transparent about AI practices also helps with regulatory compliance and can lead to cost reductions and technological innovation. These findings can guide banks in optimizing their AI-related practices to drive positive financial outcomes.

To understand AI adoption decisions within an organization, it is important to consider company culture, corporate governance, leadership, ownership structure, risk appetite, and change management. These factors shape the organization's approach to AI and impact how decisions are formulated and executed. By analyzing these dimensions, organizations can make informed decisions that align with their unique characteristics and aspirations.

Analyzing the impacts of AI implementation at the business unit or process level can reveal targeted performance improvements, enhanced decision making, efficiency gains, cost savings, customization, risk assessment, change management insights, competitive advantage, strategic alignment, and new performance metrics. This approach provides a detailed perspective on the transformational effects of AI within an organization and maximizes its potential benefits.

We identify some limitations of our research in terms of data availability; some of the annual reports are not published on the banks' websites or are not available in the English language. In addition, some annual reports are not available as PDF files but rather as scanned images, which prevents analyzing these through computerized software. Future research may also consider other banks' published information or publications of third parties rather than annual reports (e.g., bank websites, brochures, and social media advertisement tools). This paper is limited to the available data. Therefore, caution should be taken before generalizing the study's findings.

Author Contributions: Conceptualization, F.S.S., A.B.A., Q.M.O. and H.A.; methodology, F.S.S., A.B.A., Q.M.O. and H.A.; software, F.S.S., A.B.A., Q.M.O. and H.A.; validation, F.S.S., A.B.A., Q.M.O. and H.A.; formal analysis, F.S.S., A.B.A., Q.M.O. and H.A.; investigation, F.S.S., A.B.A., Q.M.O. and H.A.; resources, F.S.S., A.B.A., Q.M.O. and H.A.; data curation, F.S.S., A.B.A., Q.M.O. and H.A.; writing—original draft preparation, F.S.S., A.B.A., Q.M.O. and H.A.; writing—review and editing, F.S.S., A.B.A., Q.M.O. and H.A.; visualization, F.S.S., A.B.A., Q.M.O. and H.A.; supervision, F.S.S., A.B.A., Q.M.O. and H.A.; project administration, F.S.S., A.B.A., Q.M.O. and H.A.; funding acquisition, F.S.S., A.B.A., Q.M.O. and H.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data Available upon request.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Keywords search items and list of AI-related disclosure terms.

Column A	Column B
AI-related terms disclosure/keywords search item	
Digital transformation, fintech, financial technology, modern technology, AI digital strategy, latest technology, advanced technology, AI computing technology, digital infrastructure, digital library, electronic channel, electronic system, Internet of things, 5g technology, advanced technical experiences, digital platform, digital technology, 5g networks, blockchain, smart connection, digital awareness, digital capabilities, digital culture, digital economy, digital futuristic, digital transition, augmented reality technology, technical platform, web technology, machine learning, deep learning, augmented intelligence, natural language processing (NLP)	Digital awareness, transformation, and capabilities
Robotics, robo-advisors, automation, digital banking, mobile banking, online banking, digital services, mobile apps, electronic payment, Internet banking service, mobile branches, mobile payment, robots, mobile ATMs, digital payment, digital identity, smartphones, smart bank websites, digital product, electronic service, intelligently analyses, digital wallet, electronic wallet, mobile device service	AI application, product, service, and process
Information security, cyber security, electronic security, it risks security, electronic security policies, card security, cyber risk, electronic security, cybercrime, bank electronic security, customized electronic security methods, cyber breach, cyber resiliency, defense technology, cyber intelligence, electronic attack, global security, information security breaches, security vulnerabilities	AI information challenges and cyber security threats

Appendix B

Table A2. AI-related terms disclosure frequency related to digital awareness, transformation, and capabilities.

AI Digital Awareness, Transformation, and Capabilities	Frequency
Digital transformation	143
Fintech	130
Financial technology	85
Modern technology	13
AI	13
Digital strategy	10
Latest technology	10
Advanced technology	8
AI technology computing	4
Digital infrastructure	3
Digital library	3
Electronic channel	3
Electronic system	3

Table A2. *Cont.*

AI Digital Awareness, Transformation, and Capabilities	Frequency
Internet of things (IoT)	3
Machine learning	3
Each of these words: 5g technology, advanced technical experiences, digital platform, digital technology, 5g networks, blockchain, and smart connection	2
Digital capabilities, digital culture, digital economy, digital futuristic, digital transition, augmented reality technology, technical platform, and web technology	1
Sub-category total	457
Total words	1925
Percentage	0.24

Table A3. AI-related terms disclosure frequency related to AI applications, products, services, and processes.

AI Application, Product, Service, and Process	Frequency
Robotic automation	140
Digital banking	85
Mobile banking	80
Online banking	75
Digital services	64
Mobile app	48
Electronic payment	34
Internet banking service	20
Mobile branch	16
Mobile payment	16
Robotic/robot	16
Mobile ATM	13
Digital payment	11
Digital identity	6
Smartphone	6
Smart bank website	4
Digital products, electronic services, intelligent analyses	2
Digital wallet, electronic wallet, mobile device service	1
Sub-category total	644
Total—AI-related terms/words	1925
Percentage	0.33

Table A4. AI-related terms disclosure related to AI information challenges and cyber security threats.

AI Information Challenges and Cyber Security Threats	Frequency
Information security	523
Cyber security	138
Electronic security risks	37
IT security	33
Electronic banking services	31
Electronic security policies	24
Card security	9
Cyber risk	6
Electronic security	5
Cybercrime, bank electronic security, customized electronic security methods, cyber breach, cyber resiliency, and defense technology	2
Cyber intelligence, electronic attacks, information security breaches, and security vulnerabilities	1
Sub-total—AI-related terms/words	824
Total—AI-related terms/words	1925
Percentage	43%

Appendix C

Table A5. The OLS regressions of AI disclosure on financial performance before and during COVID-19.

	Before COVID-19					During COVID-19				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	ROA	ROE	PE	NII	TEXP	ROA	ROE	PE	NII	TEXP
AIFREC	0.00369 **	0.0266 **	−0.0237	0.0246	−0.00351 *	0.00734 **	0.0570 ***	0.143 **	0.0229	−0.00565 *
	(2.25)	(1.05)	(−0.23)	(0.26)	(−2.05)	(−2.04)	(−1.93)	(2.30)	(0.80)	(−2.12)
LASHR	−1.695 ***	−13.67 ***	23.85	−4.252 ***	−0.195	−1.140 ***	−10.66 ***	−4.701	−1.968 ***	−0.00448
	(−3.39)	(−3.19)	(1.37)	(−4.33)	(−0.57)	(−3.80)	(−4.33)	(−0.91)	(−2.89)	(−0.02)
DEBTH	−0.0397	0.127	−0.393	0.0202	−0.0621 ***	−0.0446 **	0.352 **	0.132	0.0998 **	−0.0220
	(−1.50)	(0.56)	(−0.43)	(0.39)	(−3.44)	(−2.41)	(2.32)	(0.41)	(2.37)	(−1.45)
FORSH	1.331 **	9.133 *	−20.27	0.217	0.696 *	1.452 **	14.92 ***	−13.25	0.0715	0.973 *
	(2.30)	(1.84)	(−1.00)	(0.19)	(1.76)	(2.34)	(2.93)	(−1.23)	(0.05)	(1.91)
BDSIZ	0.0417	0.145	−1.989	0.537 ***	0.0338	0.0141	0.0912	0.695	0.327 ***	−0.00483
	(0.61)	(0.25)	(−0.83)	(3.99)	(0.72)	(0.49)	(0.38)	(1.39)	(4.98)	(−0.20)
INDPB	−0.736	−6.069	31.93	1.241	−0.169	−1.380 ***	−15.17 ***	−11.18	−1.740	−0.0312
	(−1.19)	(−1.14)	(1.48)	(1.02)	(−0.40)	(−2.71)	(−3.63)	(−1.27)	(−1.50)	(−0.07)
BKSIZ	−0.296	−2.087	−5.057	−1.201 ***	0.802 ***	0.0766	−0.246	1.311	−0.309	0.154
	(−1.42)	(−1.17)	(−0.69)	(−2.92)	(5.62)	(0.58)	(−0.22)	(0.57)	(−1.02)	(1.41)
BRNCH	0.0123 ***	0.104 ***	0.0812	0.0750 ***	0.00269	0.00354	0.0624 ***	0.0193	0.0403 ***	0.0112 ***
	(2.98)	(2.94)	(0.56)	(9.25)	(0.95)	(1.25)	(2.68)	(0.39)	(6.25)	(4.81)
BKAGE	−0.0270 ***	−0.234 ***	0.255	−0.140 ***	0.00335	−0.0123 **	−0.171 ***	−0.0276	−0.0705 ***	0.00929 **
	(−3.22)	(−3.25)	(0.87)	(−8.48)	(0.58)	(−2.40)	(−4.07)	(−0.31)	(−6.05)	(2.21)
_cons	11.76 **	52.21	139.1	39.35 ***	5.314	4.716	−1.974	−29.81	14.73 **	15.35 ***
	(2.36)	(1.22)	(0.80)	(4.02)	(1.56)	(1.62)	(−0.08)	(−0.59)	(2.22)	(6.43)
N	45	45	45	45	45	82	82	82	82	82
R^2	0.490	0.518	0.207	0.851	0.951	0.452	0.489	0.215	0.678	0.891
adj. R^2	0.358	0.394	0.104	0.812	0.939	0.384	0.425	0.117	0.637	0.878

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

References

Abdallah, Abed Al-Nasser, and Ahmad K. Ismail. 2017. Corporate governance practices, ownership structure, and corporate performance in the GCC countries. *Journal of International Financial Markets, Institutions and Money* 46: 98–115. [CrossRef]

Agarwal, Shubhi. 2020. Literature Review on the Relationship between Board Structure and Firm Performance (June 30, 2020). *International Journal of Business Ethics in Developing Economies* 9: 33–43.

Albitar, Khaldoon, Khaled Hussainey, Nasir Kolade, and Ali Meftah Gerged. 2020. ESG disclosure and firm performance before and after IR. *International Journal of Accounting & Information Management* 28: 429–44. [CrossRef]

Alcácer, Vitor, and Virgilio Cruz-Machado. 2019. Scanning the Industry 4.0: A Literature Review on Technologies for Manufacturing Systems. *Engineering Science and Technology, an International Journal* 22: 899–919. [CrossRef]

Al-Gamrh, Bakr, Redhwan Al-Dhamari, Akanksha Jalan, and Asghar Afshar Jahanshahi. 2020. The impact of board independence and foreign ownership on financial and social performance of firms: Evidence from the UAE. *Journal of Applied Accounting Research* 21: 201–29. [CrossRef]

Alkaraan, Fadi, Khaldoon Albitar, Khaled Hussainey, and Vg Venkatesh. 2022. Corporate transformation toward Industry 4.0 and financial performance: The influence of environmental, social, and governance (ESG). *Technological Forecasting and Social Change* 175: 121423. [CrossRef]

Al-Malkawi, Husam-Aldin Nizar, Rekha Pillai, and M. Ishaq Bhatti. 2014. Corporate governance practices in emerging markets: The case of GCC countries. *Economic Modelling* 38: 133–41. [CrossRef]

Almustafa, Hamza, Quang Khai Nguyen, Jia Liu, and Van Cuong Dang. 2023. The impact of COVID-19 on firm risk and performance in MENA countries: Does national governance quality matter? *PLoS ONE* 18: e0281148. [CrossRef]

An, Yi, Howard Davey, and Ian R. C. Eggleton. 2011. Towards a comprehensive theoretical framework for voluntary IC disclosure. *Journal of Intellectual Capital* 12: 571–85. [CrossRef]

Anastasi, Sara, Marianna Madonna, and Luigi Monica. 2021. Implications of embedded artificial intelligence—Machine learning on safety of machinery. *Procedia Computer Science* 180: 338–43. [CrossRef]

Araújo Júnior, Jonas Ferreira, Marcelle Colares Oliveira, Vera Maria Rodrigues Ponte, and Maisa de Sousa Ribeiro. 2014. Social disclosure of Brazilian and UK firms in light of Stakeholder Theory, Legitimacy Theory and Voluntary Disclosure Theory. *Advances in Scientific and Applied Accounting* 7: 175–200. [CrossRef]

Barako, Dulacha G., Phil Hancock, and H. Y. Izan. 2006. Factors Influencing Voluntary Corporate Disclosure by Kenyan Companies. *Corporate Governance: An International Review* 14: 107–25. [CrossRef]

Berger, Allen N., and Lamont K. Black. 2011. Bank size, lending technologies, and small business finance. *Journal of Banking & Finance* 35: 724–35. [CrossRef]

Bertomeu, Jeremy, Anne Beyer, and Ronald A. Dye. 2011. Capital structure, cost of capital, and voluntary disclosures. *The Accounting Review* 86: 857–86. [CrossRef]

Bini, Stefano A. 2018. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? *The Journal of arthroplasty* 33: 2358–61. [CrossRef]

Bonsón, Enrique, and Michaela Bednárová. 2013. Corporate LinkedIn practices of Eurozone companies. *Online Information Review* 37: 969–84. [CrossRef]

Bonsón, Enrique, Domenica Lavorato, Rita Lamboglia, and Daniela Mancini. 2021. Artificial intelligence activities and ethical approaches in leading listed companies in the European Union. *International Journal of Accounting Information Systems* 43: 100535. [CrossRef]

Bonsón, Enrique, Michaela Bednárová, and David Perea. 2023. Disclosures about algorithmic decision making in the corporate reports of Western European companies. *International Journal of Accounting Information Systems* 48: 100596. [CrossRef]

Bughin, Jacques, Eric Hazan, Sreenivas Ramaswamy, Michael Chui, Tera Allas, Peter Dahlstrom, Nicolaus Henke, and Monica Trench. 2017. Artificial Intelligence: The Next Digital Frontier? McKinsey Global Institute. Available online: https://api.semanticscholar.org/CorpusID:169551387 (accessed on 10 August 2023).

Burgess, Andrew. 2017. *The Executive Guide to Artificial Intelligence: How to Identify and Implement Applications for AI in Your Organization*. Cham: Springer International Publishing. [CrossRef]

Burström, Thommie, Vinit Parida, Tom Lahti, and Joakim Wincent. 2021. AI-enabled business-model innovation and transformation in industrial ecosystems: A framework, model and outline for further research. *Journal of Business Research* 127: 85–95. [CrossRef]

Cam, Arif, Michael Chui, and Bryce Hall. 2019. Global AI survey: AI proves its worth, but few scale impact, McKinsey and Company. Available online: https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Global%20AI%20Survey%20AI%20proves%20its%20worth%20but%20few%20scale%20impact/Global-AI-Survey-AI-proves-its-worth-but-few-scale-impact.pdf (accessed on 11 April 2023).

Campbell, David, Barrie Craven, and Philip J. Shrives. 2003. Voluntary social reporting in three FTSE sectors: A comment on perception and legitimacy. *Accounting, Auditing & Accountability Journal* 16: 558–81. [CrossRef]

Cheynel, Edwige. 2013. A theory of voluntary disclosure and cost of capital. *Review of Accounting Studies* 18: 987–1020. [CrossRef]

Coad, Alex, Jacob Rubæk Holm, Jackie Krafft, and Francesco Quatraro. 2018. Firm age and performance. *Journal of Evolutionary Economics* 28: 1–11. [CrossRef]

Cockburn, Iain M., Rebecca Henderson, and Scott Stern. 2018. The impact of artificial intelligence on innovation: An exploratory analysis. In *The Economics of Artificial Intelligence: An Agenda*. Chicago: University of Chicago Press, pp. 115–46. Available online: http://www.nber.org/chapters/c14006 (accessed on 2 July 2023).

Coles, Jeffrey L., Naveen D. Daniel, and Lalitha Naveen. 2008. Boards: Does one size fit all? *Journal of Financial Economics* 87: 329–56. [CrossRef]

Couchoro, Mawuli K., Koffi Sodokin, and Moubarak Koriko. 2021. Information and communication technologies, artificial intelligence, and the fight against money laundering in Africa. *Strategic Change* 30: 281–91. [CrossRef]

Dang, Van Cuong, and Quang Khai Nguyen. 2021. Internal corporate governance and stock price crash risk: Evidence from Vietnam. *Journal of Sustainable Finance & Investment*, 1–18. [CrossRef]

Doumpos, Michael, Constantin Zopounidis, Dimitrios Gounopoulos, Emmanouil Platanakis, and Wenke Zhang. 2023. Operational research and artificial intelligence methods in banking. *European Journal of Operational Research* 306: 1–16. [CrossRef]

ElKelish, Walaa Wahid, and Mostafa Kamal Hassan. 2014. Organizational culture and corporate risk disclosure: An empirical investigation for United Arab Emirates listed companies. *International Journal of Commerce and Management* 24: 279–99. [CrossRef]
Elmarzouky, Mahmoud, Khaldoon Albitar, Atm Karim, and Ahmed Moussa. 2021. COVID-19 Disclosure: A Novel Measurement and Annual Report Uncertainty. *Journal of Risk and Financial Management* 14: 616. [CrossRef]
Eren, Berrin Arzu. 2021. Determinants of customer satisfaction in chatbot use: Evidence from a banking application in Turkey. *International Journal of Bank Marketing* 39: 294–311. [CrossRef]
European Parliament. 2023. EU AI Act: First Regulation on Artificial Intelligence. Available online: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (accessed on 17 August 2023).
Fama, Eugene F., and Michael C. Jensen. 1983. Separation of ownership and control. *Journal of Law and Economics* 26: 301–25. [CrossRef]
Fang, Hong-Xing, and Yun-Yun Jin. 2012. Listed companies' contribution to non-shareholders stakeholders: Influencing factors and value relevance of voluntary disclosure. Paper presented at 2012 International Conference on Management Science & Engineering 19th Annual Conference Proceedings, Dallas, TX, USA, September 20–22; pp. 685–92. [CrossRef]
Fatihudin, Didin. 2018. How measuring financial performance. *International Journal of Civil Engineering and Technology (IJCIET)* 9: 553–57.
Felzmann, Heike, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamó-Larrieux. 2020. Towards transparency by design for artificial intelligence. *Science and Engineering Ethics* 26: 3333–61. [CrossRef] [PubMed]
Fethi, Meryem Duygun, and Fotios Pasiouras. 2010. Assessing bank efficiency and performance with operational research and artificial intelligence techniques: A survey. *European Journal of Operational Research* 204: 189–98. [CrossRef]
Finkenwirth, Katharina. 2021. The Impact of Artificial Intelligence on Financial Performance in the German Financial Service Industry—A Content Analysis. Working Paper. Available online: https://run.unl.pt/bitstream/10362/122610/1/2020-21_fall_41614_katharina-finkenwirth.pdf (accessed on 11 April 2023).
FSB. 2017. Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications. Available online: https://www.fsb.org/wp-content/uploads/P011117.pdf (accessed on 16 May 2023).
Gander, James P. 2012. Firm debt structure, firm size and risk volatility in US industrial firms. *Applied Financial Economics* 22: 387–93. [CrossRef]
Garcia-Bedoya, Olmer, Oscar Granados, and José Cardozo Burgos. 2020. AI against money laundering networks: The Colombian case. *Journal of Money Laundering Control* 24: 49–62. [CrossRef]
Gokhale, Nikhil, Ankur Gajjaria, Rob Kaye, and Dave Kuder. 2019. AI Leaders in Financial Services, Deloitte Center for Financial Services Analysis. Available online: https://www2.deloitte.com/us/en/insights/industry/financial-services/artificial-intelligence-ai-financial-services-frontrunners.html (accessed on 8 May 2023).
Green, Lawrence W., Judith M. Ottoson, César García, and Robert A. Hiatt. 2009. Diffusion Theory and Knowledge Dissemination, Utilization, and Integration in Public Health. *Annual Review of Public Health* 30: 151–74. [CrossRef]
Haenlein, Michael, and Andreas Kaplan. 2019. A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review* 61: 5–14. [CrossRef]
Hagel, John, John Seely Brown, Tamara Samoylova, and Michael Lui. 2013. *Success or Strugle: ROA as a True Meassure of Business Performance*. New York: Deloitte University Press. Available online: http://dupress.com/articles/success-or-struggle-roa-as-a-true-measure-of-business-performance/ (accessed on 11 April 2023).
Han, Jingguang, Yuyun Huang, Sha Liu, and Kieran Towey. 2020. Artificial intelligence for anti-money laundering: A review and extension. *Digital Finance* 2: 211–39. [CrossRef]
Haninun, Haninun, Lindrianasari Lindrianasari, and Angrita Denziana. 2018. The effect of environmental performance and disclosure on financial performance. *International Journal of Trade and Global Markets* 11: 138–48. [CrossRef]
Harimaya, Kozo, and Kazumine Kondo. 2012. Determinants of Branch Expansion by Japanese Regional Banks. *The IUP Journal of Bank Management* XI: 7–25.
Hasan, Fakhrul, Mary Fiona Ross Bellenstedt, and Mohammad Raijul Islam. 2023. The impact of demand and supply disruptions during the Covid-19 crisis on firm productivity. *Global Journal of Flexible Systems Management* 24: 87–105. [CrossRef] [PubMed]
Hashem, Firas, and Rateb Alqatamin. 2021. Role of artificial intelligence in enhancing efficiency of accounting information system and non-financial performance of the manufacturing companies. *International Business Research* 14: 1–65. [CrossRef]
Hassanein, Ahmed, Alaa Mansour Zalata, and Khaled Hussainey. 2019. Do forward-looking narratives affect investors' valuation of UK FTSE all-shares firms? *Review of Quantitative Finance and Accounting* 52: 493–519. [CrossRef]
Heikal, Mohd, Muammar Khaddafi, and Ainatul Ummah. 2014. Influence Analysis of Return on Assets (ROA), Return on Equity (ROE), Net Profit Margin (NPM), Debt to Equity Ratio (DER), and current ratio (CR), Against Corporate Profit Growth in Automotive in Indonesia Stock Exchange. *The International Journal of Academic Research in Business and Social Sciences* 4: 101–14. [CrossRef]
Hirtle, Beverly. 2007. The impact of network size on bank branch performance. *Journal of Banking and Finance* 31: 3782–805. [CrossRef]
Hussainey, Khaled, Khaldoon Albitar, and Fadi Alkaraan. 2022. Corporate narrative reporting on Industry 4.0 technologies: Does governance matter? *International Journal of Accounting & Information Management* 30: 457–76. [CrossRef]
Hwang, Sewoong, and Jonghyuk Kim. 2021. Toward a Chabot for financial sustainability. *Sustainability* 13: 3173. [CrossRef]

IOSCO. 2020. The Use of Artificial Intelligence and Machine Learning by Market Intermediaries and Asset Managers: Consultation Report. The International Organization of Securities Commissions. Available online: http://www.iosco.org (accessed on 5 April 2023).

Jensen, Michael C. 1993. The modern industrial revaluation, exit, and the failure of internal control system. *Journal of Finance* 48: 831–80. [CrossRef]

Jensen, Michael C., and William H. Meckling. 1976. The economic nature of the firm: Theory of the firm: Managerial behavior, agency costs, and ownership structure. *Journal of Financial Economics* 3: 305–60. [CrossRef]

Karim, Shakir, Raj Sandu, and Mahesh Kayastha. 2021. The challenges and opportunities of adopting artificial intelligence (AI) in Jordan's healthcare transformation. *Global Journal of Information Technology: Emerging Technologies* 11: 35–46. [CrossRef]

Kaya, Orçun. 2019. Artificial Intelligence in Banking: A Lever for Profitability with Limited Implementation to Date. EU Monitor. Deutsche Bank Research. Available online: https://www.dbresearch.com/PROD/RPS_ENPROD/Artificial_intelligence_in_banking%3A_A_lever_for_pr/RPS_EN_DOC_VIEW.calias?rwnode=PROD0000000000435631&ProdCollection=PROD0000000000495172 (accessed on 23 May 2023).

Kikan, Deepak, Sumeet Singh, and Yudhvir Singh. 2019. Predictive Analytics Adoption by Banking and Financial Services: The Future Perspective. *International Journal of Recent Technology and Engineering* 8: 832–37.

Kim, Kee-Hoon, Chang-Seok Lee, Sang-Muk Jo, and Sung-Bae Cho. 2015. Predicting the success of bank telemarketing using deep convolutional neural network. Paper presented at 2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR), Fukuoka, Japan, November 13–15; pp. 314–17.

Kok, Joost N., Egbert J. W. Boers, Walter A. Kosters, Peter van der Putten, and Mannes Poel. 2002. Artificial intelligence: Definition, trends, techniques, and cases. In *Knowledge for Sustainable Development: An Insight into the Encyclopedia of Life Support Systems*. London: The UNESCO Digital Library, vol. 1, pp. 1095–107. ISBN 978-92-3-103854-9.

Kondo, Kazumine. 2018. Does branch network size influence positively the management performance of Japanese regional banks? *Applied Economics* 50: 6061–72. [CrossRef]

Krakowski, Sebastian, Johannes Luger, and Sebastian Raisch. 2022. Artificial intelligence and the changing sources of competitive advantage. *Strategic Management Journal* 44: 1425–52. [CrossRef]

Kute, Dattatray Vishnu, Biswajeet Pradhan, Nagesh Shukla, and Abdullah Alamri. 2021. Deep Learning and Explainable Artificial Intelligence Techniques Applied for Detecting Money Laundering–A Critical Review. *IEEE Access* 9: 82300–17. [CrossRef]

Lichtenthaler, Ulrich. 2020. Beyond artificial intelligence: Why companies need to go the extra step. *Journal of Business Strategy* 41: 19–26. [CrossRef]

Lightstone, Karen, and Cathy Driscoll. 2008. Disclosing elements of disclosure: A test of legitimacy theory and company ethics. *Canadian Journal of Administrative Sciences* 25: 7–21. [CrossRef]

Loderer, Claudio, René Stulz, and Urs Waelchli. 2017. Firm Rigidities and the Decline in Growth Opportunities. *Management Science* 63: 3000–20. [CrossRef]

Lu, Sylvia. 2021. Algorithmic Opacity, Private Accountability, and Corporate Social Disclosure in the Age of Artificial Intelligence. *Vanderbilt Journal of Entertainment and Technology Law* 23: 99–160.

Magness, Vanessa. 2006. Strategic posture, financial performance and environmental disclosure: An empirical test of legitimacy theory. *Accounting, Auditing & Accountability Journal* 19: 540–63. [CrossRef]

Mallinguh, Edmund, Christopher Wasike, and Zeman Zoltan. 2020. The Business Sector, Firm Age, and Performance: The Mediating Role of Foreign Ownership and Financial Leverage. *International Journal of Financial Studies* 8: 79. [CrossRef]

McKinsey & Company. 2021. Global Banking Practice: Building the AI Bank of the Future. Available online: https://www.mckinsey.com/~/media/mckinsey/industries/financial%20services/our%20insights/building%20the%20ai%20bank%20of%20the%20future/building-the-ai-bank-of-the-future.pdf (accessed on 20 July 2023).

McWaters, Jesse. 2018. The New Physics of Financial Services. In *Part of the Future of Financial Services Series: Understanding How Artificial Intelligence Is Transforming the Financial Ecosystem*. Davos: Deloitte, World Economic Forum, p. 167. Available online: http://www3.weforum.org/docs/WEF_New_Physics_of_Financial_Services.pdf (accessed on 1 August 2023).

Meiryani, Meiryani, Caineth Delvin Tandyopranoto, Jason Emanuel, A. Lindawati, Mochammad Fahlevi, Mohammed Aljuaid, and Fakhrul Hasan. 2022. The effect of global price movements on the energy sector commodity on bitcoin price movement during the COVID-19 pandemic. *Heliyon* 8: e10820. [CrossRef]

Monferrer Tirado, Diego, Moliner Tena, Miguel Ángel, and Marta Guillén Estrada. 2019. Ambidexterity as a Key Factor in Banks' Performance: A Marketing Approach. *Journal of Marketing Theory and Practice* 27: 227–50. [CrossRef]

Nguyen, Dung Minh, Yen-Ting Helena Chiu, and Huy Duc Le. 2021. Determinants of continuance intention towards banks' chatbot services in Vietnam: A necessity for sustainable development. *Sustainability* 13: 7625. [CrossRef]

Nguyen, Quang Khai. 2022. The effect of FinTech development on financial stability in an emerging market: The role of market discipline. *Research in Globalization* 5: 100105. [CrossRef]

Nguyen, Quang Khai, and Van Cuong Dang. 2023. The impact of FinTech development on stock price crash risk and the role of corporate social responsibility: Evidence from Vietnam. *Business Strategy & Development*, 1–14. [CrossRef]

Nikitas, Alexandros, Kalliopi Michalakopoulou, Eric Tchouamou Njoya, and Dimitris Karampatzakis. 2020. Artificial Intelligence, Transport and the Smart City: Definitions and Dimensions of a New Mobility Era. *Sustainability* 12: 2789. [CrossRef]

OECD. 2019. *Scoping the OECD AI Principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO)*. OECD Digital Economy Papers, No. 291. Paris: OECD Publishing. [CrossRef]

OECD. 2021. Artificial Intelligence, Machine Learning, and Big Data in Finance. Available online: https://www.oecd.org/finance/financial-markets/Artificial-intelligence-machine-learning-big-data-in-finance.pdf (accessed on 23 May 2023).

Omar, Siti Aisyah, Farhana Hasbolah, and Ulfah Mansurah Zainudin. 2017. The Diffusion of Artificial Intelligence in Governance of Public Listed Compa-nies in Malaysia. *International Journal of Business, Economics and Law* 14: 1–9.

Osburg, Thomas. 2017. Sustainability in a Digital World Needs Trust. In *Sustainability in a Digital World. CSR, Sustainability, Ethics & Governance*. Edited by Thomas Osburg and Christiane Lohrmann. Cham: Springer. [CrossRef]

Ottosson, Frida, and Martin Westling. 2020. Artificial Intelligence and its Breakthrough in the Nordics: A Study of the Relationship between AI Usage and Financial Performance in the Nordic Market (Dissertation). Available online: https://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-172725 (accessed on 15 July 2023).

Öztemel, Ercan, and Samet Gursev. 2020. Literature review of Industry 4.0 and related technologies. *Journal of Intelligent Manufacturing* 31: 127–82. [CrossRef]

Patil, Kanchan, and Mugdha Kulkarni. 2019. Artificial Intelligence in Financial Services: Customer Chatbot Advisor Adoption. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* 9: 4296–303. [CrossRef]

Perrault, Raymond, Yoav Shoham, Erik Brynjolfsson, Jack Clark, John Etchemendy, Barbara Grosz, Terah Lyons, James Manyika, Saurabh Mishra, and Juan Carlos Niebles. 2019. *The AI Index 2019 Annual Report*. Stanford: AI Index Steering Committee, Human-Centered AI Institute, Stanford University. Available online: https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf (accessed on 10 July 2023).

Plastino, Eduardo, and Mark Purdy. 2018. Game changing value from Artificial Intelligence: Eight strategies. *Strategy & Leadership* 46: 16–22. [CrossRef]

Purdy, Mark, and Paul Daugherty. 2016. *Why Artificial Intelligence is the Future of Growth*. Research Report. Dublin: Accenture Institute for High Performance, pp. 1–72. Available online: https://dl.icdst.org/pdfs/files2/2aea5d87070f0116f8aaa9f545530e47.pdf (accessed on 5 July 2023).

PwC. 2020. How Mature Is AI Adoption in Financial Services? A PwC Study across the DACH Region. Available online: https://www.pwc.de/de/future-of-finance/how-mature-is-ai-adoption-in-financialservices.pdf (accessed on 22 August 2020).

Raj, S. Benson Edwin, and A. Annie Portia. 2011. Analysis on credit card fraud detection methods. Paper presented at International Conference on Computer, Communication and Electrical Technology (ICCCET), Tirunelveli, India, March 18–19; pp. 152–56. Available online: https://api.semanticscholar.org/CorpusID:41966176 (accessed on 20 June 2023).

Rao, Anand S., and Gerard Verweij. 2017. *Sizing the Prize: What's the Real Value of AI for Your Business and How Can You Capitalize*. Jakarta: PwC Publication, pp. 1–30. Available online: https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf (accessed on 1 August 2023).

Ratia, M., Jussi Myllärniemi, and Nina Helander. 2018. Robotic process automation-creating value by digitalizing work in the private healthcare. In Paper presented at Mindtrek '18: Proceedings of the 22nd International Academic Mindtrek Conference, Tampere, Finland, October 10–11; pp. 222–27. [CrossRef]

Ryll, Lukas, Mary Emma Barton, Bryan Zheng Zhang, Jesse McWaters, Emmanuel Schizas, Rui Hao, Keith Bear, Massimo Preziuso, Elizabeth Seger, and Robert Wardrop. 2020. Transforming Paradigms: A Global AI in Financial Services Survey. Electronic. World Economic Forum. Available online: https://www.weforum.org/reports/transforming-paradigms-a-global-ai-in-financialservices-survey (accessed on 20 June 2023).

Saenz, Maria Jesus, Elena Revilla, and Cristina Simón. 2020. Designing AI Systems with Human-Machine Teams. *MIT Sloan Management Review* 61: 1–5.

Sætra, Henrik Skaug. 2021. A Framework for Evaluating and Disclosing the ESG Related Impacts of AI with the SDGs. *Sustainability* 13: 8503. [CrossRef]

Sanny, Lim, Ari Clementin Susastra, Choky Roberts, and Reflia Yusramdaleni. 2020. The analysis of customer satisfaction factors which influence Chabot acceptance in Indonesia. *Management Science Letters* 10: 1225–32. [CrossRef]

Shang, Zhiming, and Liming Zhang. 2022. The Sustainable Digitalization in the Manufacturing Industry: A Bibliometric Analysis and Research Trend. *Mobile Information Systems* 2022: 1–11. [CrossRef]

Shehata, Nermeen Fathy. 2014. Theories and Determinants of Voluntary Disclosure. *Accounting and Finance Research* 3: 18–26. [CrossRef]

Shen, Huayu, Mengyao Fu, Hongyu Pan, Zhongfu Yu, and Yongquan Chen. 2020. The impact of the COVID-19 pandemic on firm performance. *Emerging Markets Finance and Trade* 56: 2213–30. [CrossRef]

Shiyyab, Fadi Shehab, Claudia Girardone, and Idlan Zakaria. 2014. Pay for no performance? Executive pay and performance in EU banks. Paper presented at European Financial Management Association 2014 Annual Meetings, Rome, Italy, June 25–28; Available online: https://www.efmaefm.org/0efmameetings/EFMA%20ANNUAL%20MEETINGS/2014-Rome/accepted%20papers/Z.php (accessed on 25 August 2023).

Shook, Ellyn, and Mark Knickrehm. 2018. *Future Workforce Survey—Banking Realizing the Full Value of AI*. Dublin: Accenture. Available online: https://epicdisruption.accenture.com/content/pdf/Accenture-Reworking-the-Revolution.pdf (accessed on 22 August 2023).

Soni, Vishal D. 2019. Role of Artificial Intelligence in Combating Cyber Threats in Banking. *International Engineering Journal for Research & Development* 4: 7. [CrossRef]

Suhel, Sasha F., Vinod S. Shukla, Sonali Vyas, and Ved P. Mishra. 2020. Conversation to Automation in Banking Through Chatbot Using Artificial Machine Intelligence Language. Paper presented at 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, June 4–5; pp. 611–18. Available online: https://ieeexplore.ieee.org/document/9197825 (accessed on 11 August 2023).

Tákacs, Árpad, Imre Rudas, Dominik Bösl, and Tamas Haidegger. 2018. Highly Automated Vehicles and Self-Driving Cars [Industry Tutorial]. *IEEE Robotics & Automation Magazine* 25: 106–12. [CrossRef]

Tekic, Zeljko, and Dmitry Koroteev. 2019. From disruptively digital to proudly analog: A holistic typology of digital transformation strate-gies. *Business Horizons* 62: 683–93. [CrossRef]

Thorun, Christian. 2018. Corporate Digital Responsibility: Unternehmerische Verantwortung in der digitalen Welt. *In Fallstudien zur Digitalen Transformation. Edited by Christian Gärtner and Christian Heinrich. Wiesbaden: Springer Gabler*. [CrossRef]

Vijai, C., S. M. Suriyalakshmi, and Masilamani Elayaraja. 2020. The Future of Robotic Process Automation (RPA) in the Banking Sector for Better Customer Experience. *Shanlax International Journal of Commerce* 8: 61–65. [CrossRef]

Wamba-Taguimdje, Serge-Lopez, Samuel Fosso Wamba, Jean Robert Kala Kamdjoug, and Chris Emmanuel Tchatchouang Wanko. 2020. Influence of artificial intelligence (AI) on firm performance: The business value of AI-based transformation projects. *Business Process Management Journal* 26: 1893–924. [CrossRef]

Wheeler, Diamond. 2020. Coopting Artificial Intelligence as an Opportunity for Financial Service Professionals. *Journal of Financial Service Professionals* 74: 66–72.

Yermack, David. 1996. Higher market valuation of companies with a small board of directors. *Journal of Financial Economics* 40: 185–211. [CrossRef]

Yu, Hui-Cheng, Lopin Kuo, and Mao-Feng Kao. 2017. The relationship between CSR disclosure and competitive advantage. *Sustainability Accounting, Management and Policy Journal* 8: 547–70. [CrossRef]

Zakaryazad, Ashkan, and Ekrem Duman. 2016. A profit-driven Artificial Neural Network (ANN) with applications to fraud detection and direct marketing. *Neurocomputing* 175: 121–31. [CrossRef]

Zardkoohi, Asghar, Eugene Kang, Donald Fraser, and Albert A. Cannella. 2018. Managerial Risk-Taking Behavior: A Too-Big-To-Fail Story. *Journal of Business Ethics* 149: 221–33. [CrossRef]

Zetzsche, Dirk Andreas, Douglas W. Arner, Ross P. Buckley, and Brian Tang. 2020. Artificial Intelligence in Finance: Putting the Human in the Loop. CFTE Academic Paper Series: Centre for Finance, Technology and Entrepreneurship, no. 1. University of Hong Kong Faculty of Law Research Paper No. 2020/006. Available online: https://ssrn.com/abstract=3531711 (accessed on 10 August 2023).

Zhou, Shan, Roger Simnett, and Wendy Green. 2017. Does integrated reporting matter to the capital market? *Abacus* 53: 94–132. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AI

MDPI

Article

Optimized Financial Planning: Integrating Individual and Cooperative Budgeting Models with LLM Recommendations

I. de Zarzà [1,2,3,*], J. de Curtò [1,2,3], Gemma Roig [1,4] and Carlos T. Calafate [2]

[1] Informatik und Mathematik, GOETHE-University Frankfurt am Main, 60323 Frankfurt am Main, Germany; decurto@em.uni-frankfurt.de (J.d.C.); roig@cs.uni-frankfurt.de (G.R.)
[2] Departamento de Informática de Sistemas y Computadores, Universitat Politècnica de València, 46022 València, Spain; calafate@disca.upv.es
[3] Estudis d'Informàtica, Multimèdia i Telecomunicació, Universitat Oberta de Catalunya, 08018 Barcelona, Spain
[4] HESSIAN Center for AI (hessian.AI), 64293 Darmstadt, Germany
* Correspondence: dezarza@em.uni-frankfurt.de

Abstract: In today's complex economic environment, individuals and households alike grapple with the challenge of financial planning. This paper introduces novel methodologies for both individual and cooperative (household) financial budgeting. We firstly propose an optimization framework for individual budget allocation, aiming to maximize savings by efficiently distributing monthly income among various expense categories. We then extend this model to households, wherein the complexity of handling multiple incomes and shared expenses is addressed. The cooperative model prioritizes not only maximized savings but also the preferences and needs of each member, fostering a harmonious financial environment, whether they are short-term needs or long-term aspirations. A notable innovation in our approach is the integration of recommendations from a large language model (LLM). Given its vast training data and potent inferential capabilities, the LLM provides initial feasible solutions to our optimization problems, acting as a guiding beacon for individuals and households unfamiliar with the nuances of financial planning. Our preliminary results indicate that the LLM-recommended solutions result in budget plans that are both economically sound, meaning that they are consistent with established financial management principles and promote fiscal resilience and stability, and aligned with the financial goals and preferences of the concerned parties. This integration of AI-driven recommendations with econometric models, as an instantiation of an extended coevolutionary (EC) theory, paves the way for a new era in financial planning, making it more accessible and effective for a wider audience, as we propose an example of a new theory in economics where human behavior can be greatly influenced by AI agents.

Keywords: financial planning; household finance; LLMs; budgeting; extended coevolutionary

1. Introduction

Financial planning [1,2] is an indispensable part of modern life. Whether it is an individual planning their monthly expenses or a family juggling with combined finances, effective budgeting is crucial for achieving long-term financial security and prosperity. However, with the proliferation of expenses and the intricate nature of contemporary economic systems, the endeavor of financial planning often feels overwhelming to many. Traditional methods, while systematic, sometimes fall short in accommodating the dynamic nature of financial needs, and often lack the ability to personalize recommendations to cater to individual or household-specific priorities.

We are now entering the age of artificial intelligence (AI) [3], a transformative force that has already revolutionized numerous sectors, from healthcare to entertainment. The vast inferential and predictive capabilities of AI, especially models like the large language

model (LLM) [4], offer a unique advantage in the realm of financial planning. By processing vast amounts of financial data and patterns [5], such models can provide tailored recommendations that not only adhere to general financial best practices, which are widely accepted principles or guidelines in financial planning that are considered effective for managing personal finances, but are also congruent with specific user inputs.

In this paper, we set out to bridge the traditional econometric methods of financial planning with the power of AI-driven recommendations. By considering the LLM's suggestions as a foundation , we refine and optimize budget allocations for both individuals and households. Our approach strives to achieve a balance between mathematical optimization and real-world practicality, ensuring that the solutions are not just economically sound but also feasible and user-friendly.

In practical terms, when artificial neural networks (ANNs), including DL and LLMs, are trained through supervised learning, they employ a learning rule that iteratively converges toward an equilibrium, finely balancing the dilemma of underfitting versus overfitting. Despite their ability to discern nonlinear continuous transformations, their operational mechanics do not inherently encapsulate an internal representation of the nonlinear relationships between training exemplars [6]. The consequence of this limitation is twofold: firstly, when ANNs process data with inter-relations confined within Euclidean boundaries, they excel; secondly, when confronted with data existing beyond these confines, the results may not only be suboptimal but also unpredictably aberrant. This limitation is similar, in many respects, to the constraints encountered in traditional statistical regression methods.

In the subsequent sections, we will delve into the methodologies that we have developed for individual and cooperative financial planning, elucidate on the integration of LLMs in these methods, and present the results of our preliminary tests. Through this exploration, we hope to usher in a new paradigm in financial planning, one that is more accessible, efficient, and in tune with the unique needs of today's diverse society. That being said, it is also necessary to recognize the potential limitations and inherent risks of AI, such as errors and hallucinations, and the strategies that can be implemented to mitigate these issues, such as rigorous validation processes, the constant monitoring of AI outputs, and the incorporation of fail-safes and human oversight.

The essence of this study lies in addressing a critical challenge: the democratization of financial planning in a landscape riddled with complexity and unpredictability. Our work is driven by the imperative to equip individuals and households with the tools to navigate the economic maze toward sustainable financial health. The problem statement is twofold—firstly, the need to simplify the intricacies of financial decision making for the non-expert and, secondly, to create an adaptive system that is resilient to the volatile nature of economic dynamics.

Our motivation is rooted in the recognition that financial security and literacy are cornerstones of individual autonomy and societal well-being. We propose an innovative solution that intersects the predictive acuity of LLMs with the foundational constructs of econometrics. The mathematical core of our approach is to optimize a utility function U that encapsulates the financial health of an entity, of the form:

$$U(I, E) = \alpha \log(S) - \beta \sum_{o=1}^{n} w_o \log(E_o) \tag{1}$$

where I is income, S is savings, E represents expenses across n categories, w_o are the personalized preference weights, and α, β are parameters calibrating the trade-off between savings and expenditures. Through this formulation, we seek to craft a system that not only provides bespoke financial advice but also evolves with the user's changing economic scenario, embodying the fluidity required for real-world financial planning.

The problem at the heart of our research is the incremental complexity of personal financial management in a rapidly evolving economic landscape. In an age where financial instruments and markets are increasingly intricate, the average individual is at a significant disadvantage when attempting to plan for financial security. This challenge is not merely

academic but a pressing real-world issue that affects the quality of life and long-term well-being of individuals and families. The need for accessible financial planning is further highlighted by the widening gap in financial literacy, which can lead to poor financial decisions and increased economic inequality.

Our work is motivated by the imperative to render financial planning not only more approachable but also more aligned with the diverse and dynamic financial goals of individuals. We aim to address this by leveraging the computational prowess of LLMs to distill complex financial data into actionable insights. The mathematical underpinning of our approach is an optimization framework, tailored to the nuances of personal finance, which can be represented as:

$$\max_{\mathbf{x}} \{U(\mathbf{x}) : \mathbf{x} \in \mathcal{F},\ \mathcal{G}(\mathbf{x}) \leq 0,\ \mathcal{H}(\mathbf{x}) = 0\} \tag{2}$$

Here, $\mathbf{x}$ symbolizes the financial decision variables, U is the utility function that quantifies financial health, $\mathcal{F}$ defines the feasible set based on financial constraints, $\mathcal{G}$ represents inequality constraints (such as budget limits), and $\mathcal{H}$ captures equality constraints (such as income–expense balance). Our motivation is to implement a system that empowers users with personalized, intelligent recommendations, thus equipping them to make informed financial choices in pursuit of economic resilience and prosperity.

The remainder of this paper is structured as follows: Section 2 provides related work. Section 3 focuses on the materials and methodology adopted, presenting the foundational aspects and research methodology pertinent to the optimization models. In Section 4, we explore the significance of LLMs in individual budgeting, emphasizing their recommendation-driven approach to financial planning. Section 5 offers an in-depth look into cooperative financial planning for households, showcasing the synergistic view between LLM recommendations and household members' financial priorities. The subsequent Section 6 presents the intricacies of cooperative budgeting within an established extended coevolutionary (EC) framework, elucidating its innovative facets. Moving forward, Section 7 describes our simulation framework, detailing how LLM-informed decisions can be simulated and analyzed, while also highlighting in Section 7.3 the importance of long-term planning dependencies to further develop the simulated framework. Finally, in Section 8, we draw our conclusions, recapitulating the primary findings and looking ahead at promising directions for future work in this domain.

2. Related Work

The landscape of financial planning and optimization has been deeply enriched by the intersection of computational intelligence and economic theories. In this section, we contextualize our contributions within the spectrum of existing literature, delineating the evolution of financial planning methodologies from traditional models to AI-augmented frameworks.

Traditional financial planning has been heavily influenced by seminal works such as those of Markowitz [7] on portfolio optimization and Ando and Modigliani's life-cycle hypothesis [8]. The optimization of financial assets has been extensively studied, with methods ranging from linear programming [9] to more sophisticated stochastic models [10].

The infusion of behavioral economics into financial optimization models has provided a more thorough understanding of human financial behavior, as evidenced by the works of Thaler and others [11]. This has led to models that account for irrational decision making and heuristics used by individuals, adding complexity to the optimization problems [12].

The advent of AI [13,14] in financial planning has opened new avenues for personalized and adaptive financial advice. Early systems utilized rule-based expert systems [15], while recent advancements have leveraged machine learning for predictive analytics in finance [16].

The integration of LLMs [4] into financial planning is a relatively new and emerging field. Recent studies have begun to explore the capabilities of models like OpenAI's GPT

series and explainable AI techniques in generating financial narratives [17,18]. Our work extends this line of research by embedding LLM insights into the optimization framework, a novel approach that has not been extensively covered in the existing literature.

Comparative studies have highlighted the strengths and weaknesses of AI-driven financial tools compared to traditional methods [19–21]. However, there remains a gap in the literature regarding the efficacy of LLM recommendations within the cooperative budgeting context, an area our research directly addresses.

While existing works provide a foundational understanding of the financial optimization landscape, there remains a lacuna when it comes to the integration of real-time, AI-driven, personalized recommendations into these models. Our research fills this gap by not only integrating but also critically examining and refining the role of LLMs within financial planning and optimization frameworks.

In summary, our research builds upon the rich heritage of financial optimization while branching into the relatively uncharted territory of LLM-augmented financial planning. By doing so, we aim to contribute a significant advancement in the domain, tailored for the dynamic economic world of the 21st century.

3. Materials and Methodology

This research proposes an optimization model for individual and cooperative financial planning, both enhanced by the integration of large language model (LLM) recommendations. The two primary facets of this approach are: (1) foundational mathematical models that describe budgeting optimization problems, and (2) computational algorithms that incorporate the LLM's recommendations into financial decision-making processes.

The objective of maximizing savings by allocating monthly income across various expense categories is grounded in core principles of theory in household finance. This approach aligns with the life-cycle hypothesis of Ando and Modigliani (1963) [8], which posits that individuals plan their consumption and savings behavior over their lifetime to achieve financial stability. Furthermore, it reflects the principles of optimal portfolio selection from the theory of modern portfolio of Markowitz (1952) [7], adapted to the context of household budgeting, where the portfolio is a mix of expenses and savings.

Our model also incorporates the behavioral aspects of financial decision making as discussed by Xiao (2008) [22], acknowledging that individuals have varying preferences and subjective weights on different expense categories. Additionally, Browning and Lusardi's (1996) research on micro theories and micro facts of household saving [23] provides empirical support for the relevance of optimizing savings behavior at the household level.

For the cooperative financial recommendation model, we draw on these individual financial behaviors and extend them to a household context, considering the collective utility maximization and shared financial responsibilities, resonating with the cooperative game theory in household economics [24].

Individual Financial Recommendation Model

Objective: Our primary goal is to find the best allocation of an individual's monthly income across various expense categories while maximizing their potential savings [7,8,22,23].

Variables:

- I : Monthly income.
- E_z : Expense for category z (e.g., rent, groceries, utility bills, entertainment).
- S : Savings for the month.

Objective Function: This function emphasizes maximizing savings after covering all necessary expenses.

$$\text{Maximize } S = I - \sum_{z} E_z. \tag{3}$$

Constraints: The constraints ensure non-negative expenses, savings, and that total expenses do not surpass the monthly income.

$$E_z \geq 0 \, \forall z, \tag{4}$$

$$S \geq 0, \tag{5}$$

$$\sum_z E_z \leq I. \tag{6}$$

Cooperative Financial Recommendation Model (Household)

Objective:

In a household setting, our aim is to optimize combined budget allocation for all members. We ensure that each member's preferences and necessities are accommodated, aiming for an overall maximization of household savings.

Variables:

- I_o : Monthly income of member o.
- E_{zo} : Expense for category z by member o.
- S_o : Savings for the month by member o.
- T_S : Total combined savings of the household.

Objective Function:

This function prioritizes the overall savings of the household, given each member's income and expenses.

$$\text{Maximize } T_S = \sum_o (I_o - \sum_z E_{zo}). \tag{7}$$

Constraints:

The constraints are extensions of the individual model but factor in multiple members.

$$E_{zo} \geq 0 \, \forall z, o, \tag{8}$$

$$S_o \geq 0 \, \forall o, \tag{9}$$

$$\sum_z E_{zo} \leq I_o \, \forall o, \tag{10}$$

$$T_S = \sum_o S_o. \tag{11}$$

Preferences/Weights:

To personalize the model, we introduce weights w_{zo} representing the importance of an expense category z for member o. This ensures that individual preferences are not lost in the collective model.

$$\text{Maximize } T_S = \sum_o (I_o - \sum_z w_{zo} E_{zo}). \tag{12}$$

Collaborative Constraint:

Factoring in shared expenses, such as rent, can lead to substantial savings. Therefore, we introduce a combined expense constraint.

$$E_{\text{rent, combined}} \leq E_{\text{rent},1} + E_{\text{rent},2}. \tag{13}$$

Incorporating LLM Recommendations

The true innovation of our approach lies in blending traditional optimization techniques with AI-driven insights [25,26]. The LLM's vast knowledge base and advanced language processing capabilities can provide nuanced, context-aware recommendations for individuals and households.

Given a recommendation R_z from the LLM for an expense category z, we can adjust our initial financial allocations.

$$E_z \approx R_z \, \forall z. \tag{14}$$

With these LLM-generated insights serving as a foundation, the optimization model can be fine-tuned at a later stage to suit specific needs, balancing mathematical accuracy with real-world financial insights.

3.1. Prospective Validation Framework for AI-Driven LLM Financial Recommendations

This framework can be designed to uphold the integrity and applicability of AI-driven insights, ensuring that they are not only relevant but also conform to the highest standards of financial prudence.

A multi-tiered verification process that scrutinizes the LLM output for reliability and truthfulness can be proposed. This process would involve:

- Expert Review: Each LLM-generated recommendation is evaluated by financial experts to ensure that it is consistent with established economic principles and real-world conditions. Experts may modify LLM suggestions to correct any inaccuracies or to better align with conservative financial practices.
- Contextual Analysis: The model contextualizes LLM responses by examining them against current financial trends, regulatory guidelines, and economic indicators. This ensures that the advice is not only mathematically sound but also practical, taking into account the socio-economic environment of the users.
- Risk Assessment: Risk analysis is conducted to evaluate the potential harm that could arise from following the LLM's advice. This includes stress testing the recommendations against various economic scenarios to gauge their resilience and adaptability.
- Ethical Considerations: An ethical review is performed to ensure that the LLM's advice does not inadvertently lead to financial decisions that could be detrimental to the user's long-term well-being or that could perpetuate biases.

The objective function within the optimization framework would be then defined as follows:

$$\text{Maximize } U(S, \mathbf{E}, C) = f(S) - g(\mathbf{E}, C) \tag{15}$$

where $U(S, \mathbf{E}, C)$ is the utility function dependent on savings S, a vector of expenses $\mathbf{E}$, and a context vector C that includes both economic conditions and user-specific circumstances. The function $f(S)$ quantifies the utility derived from savings, emphasizing the goal of maximization, while $g(\mathbf{E}, C)$ is a penalty function that increases when expenses are either economically unsound or potentially harmful.

To integrate LLM insights effectively, a constraint adaptation strategy can be employed where LLM recommendations adjust the boundaries of the feasible solution space rather than dictating fixed allocations. This would allows us to maintain a flexible yet controlled optimization environment where LLM inputs serve as informed guidelines rather than rigid directives. The optimization problem would then become:

$$\text{Maximize } U(S, \mathbf{E}, C) \tag{16}$$

$$\text{subject to } \mathbf{E}_{\text{min}} \leq \mathbf{E} \leq \mathbf{E}_{\text{LLM}}, \tag{17}$$

$$S + \sum \mathbf{E} \leq I, \tag{18}$$

$$S, \mathbf{E} \geq 0, \tag{19}$$

where $\mathbf{E}_{\text{min}}$ and $\mathbf{E}_{\text{LLM}}$ are vectors of minimum and LLM-recommended expenses, respectively.

Through this meticulous approach, we can ensure that the financial planning methodologies proposed in this study are not only innovative and AI-enhanced but are also reliable, ethically responsible, and practically applicable in a variety of financial contexts. This methodology, with the proposal of a thorough validation mechanism, sets a precedent for the responsible integration of AI in the sensitive realm of personal finance.

3.2. RAG Approach with LLM Recommendations

Additionally, to address the potential for LLM hallucination, the methodology could integrate a retrieval augmented generation (RAG) approach into the recommendation generation process. This sophisticated mechanism can enhance the robustness of the LLM's outputs by anchoring its responses in verifiable data, thereby significantly reducing the likelihood of generating spurious or misleading financial advice.

The RAG process can operate by first querying a large-scale dataset of financial information, regulatory guidelines, and historical economic data. The LLM then synthesizes this retrieved information with its pre-trained knowledge to generate recommendations that are not only contextually rich but also data-grounded. Here is how the RAG methodology could be incorporated:

- Data Retrieval: Prior to recommendation generation, the LLM queries a curated dataset to retrieve relevant financial data and historical trends that mirror the user's query context.
- Augmented Generation: The model combines its own language-understanding capabilities with the retrieved data to generate a draft recommendation. This step ensures that each piece of advice is informed by empirical data, thereby enhancing its accuracy and relevance.
- Validation and Refinement: The draft recommendation is then passed through a validation filter, which includes a comparison against up-to-date financial models and a check for compliance with current economic conditions.
- Expert Oversight: The final stage can involve human oversight, where financial experts review the LLM-generated advice to ensure that it adheres to professional standards and reflects sound financial judgment.

The incorporation of RAG into our recommendation generation process can be mathematically represented in the optimization model as an additional set of constraints that the LLM's responses must satisfy:

$$RAG(\mathbf{E}, \mathbf{D}) \rightarrow \mathbf{E}_{\text{LLM}} \tag{20}$$

where $RAG(\mathbf{E}, \mathbf{D})$ represents the RAG process with expenses $\mathbf{E}$ and retrieved data $\mathbf{D}$, mapping to the LLM's recommended expenses $\mathbf{E}_{\text{LLM}}$.

Incorporating the RAG methodology into the LLM framework could significantly mitigate the risk of hallucinations by ensuring that the model's recommendations are continuously grounded in empirical evidence. It can also provide an additional layer of confidence in the model's outputs, reinforcing their reliability for end-users seeking sound financial planning advice.

4. LLM-Informed Budgeting

In the modern era of data-driven solutions, using an LLM to provide financial advice offers an innovative approach to household budgeting. Given the intricate nature of individual spending habits and the myriad of ways in which funds can be allocated, using the power of AI can significantly enhance personalized financial recommendations.

The value addition of our research to the field of financial planning lies in the novel synthesis of econometric modeling and state-of-the-art AI. By integrating the insights from LLMs into traditional financial planning frameworks, we provide a multi-faceted approach that aligns with the contemporary demand for personalized financial advice. This integration heralds a significant shift from one-size-fits-all financial advisories to solutions that respect individual financial narratives and the unique economic contexts that they exist in.

Our models offer significant improvements over existing methodologies by incorporating behavioral economic factors and real-time data analysis, thereby delivering a more granular and responsive financial planning tool. The mathematical model that encapsulates our approach can be represented as:

$$\max_{\mathbf{x}} \; U(\mathbf{x}, \boldsymbol{\theta}) \quad \text{subject to} \quad \mathbf{x} \in \mathcal{F}(\boldsymbol{\Psi}),\ \mathcal{G}(\mathbf{x}, \boldsymbol{\Phi}) \leq 0,\ \mathcal{H}(\mathbf{x}, \boldsymbol{\Lambda}) = 0 \tag{21}$$

Here, $\mathbf{x}$ denotes the vector of decision variables for financial planning, U is the utility function enhanced by LLM insights $\boldsymbol{\theta}$, $\mathcal{F}$ represents the feasible region influenced by external data $\boldsymbol{\Psi}$, and $\mathcal{G}$, $\mathcal{H}$ are the sets of constraints parameterized by $\boldsymbol{\Phi}$ and $\boldsymbol{\Lambda}$, respectively, signifying economic conditions and user preferences. The LLM's capability to digest and interpret large volumes of data transcends the capabilities of traditional methods, thus enabling our model to adaptively recalibrate recommendations in accordance with the fluctuating economic indicators and personal financial goals.

Moreover, the transformative potential of our methodology becomes particularly salient when considering the adaptability of LLMs to a changing information landscape, a feature that conventional financial planning tools lack. By continually learning from new data, our model remains at the forefront of financial planning innovation, marking a decisive step toward a more intelligent, adaptable, and personalized financial planning paradigm.

The approach leverages the capabilities of OpenAI's GPT-4 model to analyze a user's financial data and provide a bespoke recommendation based on that data. The process involves:

1. Data Input: Users provide a breakdown of their monthly income and primary expenses.
2. LLM Analysis: The model evaluates the data in the context of sound financial planning.
3. Recommendation Generation: The LLM then offers a detailed recommendation on how the user might best allocate their remaining funds.

The Python script, `get_financial_recommendation`, takes a string representation of the user's financial data and attempts to retrieve or generate a recommendation.

Key features of the script include:

- Caching Mechanism: To enhance efficiency, the script maintains a `recommendation_cache`. If the user's financial data match a previous entry, a stored recommendation is returned without invoking the LLM, thereby saving computational resources and time.
- Robust Error Handling: Recognizing the potential for API errors or timeouts, the script employs a retry mechanism with exponential backoff. This ensures that transient issues do not lead to immediate failures and offers a resilient approach.
- User-Friendly Output: The LLM's response is parsed to retrieve a clear, concise recommendation that is then returned to the user.

Here, we describe an example for clarity. Consider a user with a monthly income of USD 5000 and major expenses that include USD 1200 for rent, USD 400 for groceries, USD 200 for utilities, and USD 100 for entertainment. When they input these data, the LLM could provide a recommendation like:

> "Based on your financial data for the month, after accounting for your major expenses, you have $2100 remaining. I recommend allocating $500 towards an emergency fund, $300 towards investments, $200 for miscellaneous expenses, and saving the remaining $1100."

Advantages of the LLM Approach:

1. Personalization: Unlike generic financial advice, the LLM approach tailors recommendations to the individual's unique financial situation.
2. Scalability: The approach can cater to a vast number of users simultaneously, making it suitable for integration into financial planning apps or platforms.
3. Learning Over Time: As LLMs evolve and receive more training, the quality and depth of their recommendations can only improve.

While the current implementation provides robust financial recommendations, future versions could:

- Integrate real-time data sources to provide up-to-the-minute advice.
- Consider longer-term financial goals, such as buying a home or retirement planning, in its recommendations.
- Incorporate user feedback to continually refine and enhance recommendations.

In conclusion, using LLMs for household financial recommendations represents a significant step forward in the marriage of AI and financial planning. By offering personalized, scalable, and increasingly sophisticated advice, this approach promises to revolutionize the way individuals manage their finances.

Figure 1 delineates the proposed financial planning methodology that makes use of the capabilities of LLMs to enhance data analysis and supports the formulation of robust recommendations. This schematic presents a sequence of interconnected stages starting with the aggregation of financial data, proceeding to an in-depth analysis phase, which converge into the application of LLM insights for the generation of tailored financial recommendations. The iterative cycle is fundamental for maintaining the relevancy and efficacy of the financial plan, adapting dynamically to the evolving economic landscape and individual and cooperative financial objectives.

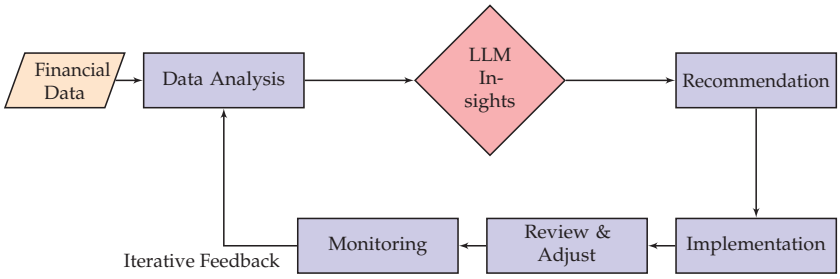


Figure 1. Flowchart illustrating the financial planning methodology with the integration of LLMs. The process starts with data collection, followed by analysis, integration of LLM insights, recommendation, implementation, and iterative review with monitoring to adjust the plan.

4.1. Verification Mechanism for LLM-Generated Financial Advice

In our approach, we could ensure the alignment of the LLM-generated recommendations with established financial truths by implementing a verification mechanism within the Python script: get_financial_recommendation . This mechanism can cross-reference the recommendations against a comprehensive financial database that includes a wide array of economic indicators, historical financial data, and regulatory compliance requirements. Here, we describe the process in detail:

- Recommendation Generation: The script first invokes the LLM to generate a preliminary financial recommendation based on the user's input data.
- Database Cross-Reference: Upon obtaining the LLM's output, the script automatically queries a financial database, seeking corroboration for the proposed financial strategies. This database is consistently updated to reflect the latest financial truths, including market trends, tax laws, and industry standards.
- Truth Alignment Check: The script compares the LLM's recommendation against the database entries to verify the accuracy and relevance of the advice. Recommendations that significantly diverge from the financial truths are flagged for further review.
- Feedback Loop for Adjustment: In cases where discrepancies arise, the script adjusts the recommendations by applying a set of heuristic rules derived from the financial truths. This iterative process ensures that the final advice provided to the user is both data-driven and aligned with financial realities.
- Human Expert Review: The last step involves a financial expert who reviews and, if necessary, refines the script's output to ensure full compliance with financial best practices.

Mathematically, this verification mechanism can be represented as a constraint satisfaction problem, where the generated recommendations R must satisfy a set of conditions C based on the truths T from the financial database D:

$$\forall r \in R, \exists t \in T : C(r, t) \rightarrow \text{Valid Recommendation.} \tag{22}$$

Incorporating this truth-checking protocol could enhance the reliability of our LLM-informed budgeting system, ensuring that users receive recommendations that are not only personalized but also pragmatically sound and financially valid.

4.2. Prompt Engineering for Improved LLM Financial Recommendations

To empower users in optimizing their interactions with the LLM, we can also introduce a guided prompt optimization feature within our LLM-informed budgeting system. This enhancement can educate users on constructing effective prompts, thereby improving the quality of the financial recommendations that they receive. Here is how we can incorporate this guidance into the user experience:

- Prompt Templates: We provide users with a collection of prompt templates that are designed based on successful financial inquiry patterns. These templates serve as starting points for users to articulate their financial data and questions.
- Prompt Feedback Mechanism: After users submit their prompts, our system analyzes the prompt structure and provides instant feedback on how to refine it for clearer communication and better-aligned financial recommendations.
- Adaptive Learning from User Interaction: As users interact with the system, it can employ ML algorithms to learn from the quality of recommendations generated from past prompts. The system then suggests modifications to the prompts for future queries, aiming to maximize the relevance and precision of the recommendations.
- Informative Tutorials and Tips: Alongside direct feedback, users have access to educational materials that explain the principles of effective prompt construction, including clarity, specificity, and the inclusion of relevant financial context.
- Expert System Insights: A built-in expert system analyzes user prompts and compares them with a database of optimized prompts and their outcomes. By highlighting key differences, the system offers actionable insights for users to improve their queries.

Mathematically, this process can be viewed as an optimization problem where the aim is to maximize the utility U of a prompt P with respect to the quality Q of the recommendation R as follows:

$$\max_{P} U(P) \quad \text{s.t.} \quad Q(R|P) \text{ is maximized.} \tag{23}$$

The function $U(P)$ denotes the utility of the prompt, which is a function of its effectiveness in generating high-quality recommendations $Q(R|P)$. By iteratively refining P based on system feedback and learning, we can ensure a continuous improvement in the prompt's utility.

4.3. EC Theory for Household Finance: Beyond Euclidean Domains

The extended coevolutionary (EC) theory provides a sophisticated framework that allows us to navigate beyond the intrinsic limitations of ANNs. While these models might struggle with capturing the intricacies of non-Euclidean spaces due to their reliance on linear transformations within a Euclidean framework, EC theory embraces a more holistic view.

EC theory, through its foundational principles of coadaptation and coevolution, inherently accounts for the complex and dynamic inter-relationships characteristic of real-world scenarios, which often extend beyond linearly definable spaces. This adaptability is made possible by EC theory's capacity to model interactions within a multi-dimensional space

that can effectively represent the high-level abstractions and nonlinear interdependencies of financial data.

$$\Delta \mathbf{x}(t+1) = F(\mathbf{x}(t), \mathbf{x}(t-1), \ldots, \mathbf{x}(0)). \tag{24}$$

Here, $\Delta \mathbf{x}(t+1)$ represents the change in the strategy vector at time $t+1$, and F encapsulates a nonlinear function of the strategy vectors at all previous time steps, capturing the essence of iterative learning and strategic adaptation over time. This function F is not restricted to the confines of a Euclidean space, enabling the EC model to explore a more diverse solution space and offering a robust mechanism for identifying equilibria in complex systems.

Furthermore, the EC theory facilitates the integration of external insights, such as those derived from LLMs, into its evolutionary process. By considering the LLM outputs as auxiliary data points, the EC model can refine its strategic outlook, effectively bridging the gap between the traditional optimization techniques and the high-dimensional space of AI-driven recommendations.

$$\mathbf{x}(t+1) = \mathbf{x}(t) + \gamma(\mathbf{R} - \mathbf{x}(t)) + \Delta \mathbf{x}(t+1). \tag{25}$$

In the equation above, γ represents the rate of integration of LLM recommendations denoted by $\mathbf{R}$ into the current strategy $\mathbf{x}(t)$. The term $\Delta \mathbf{x}(t+1)$, as previously defined, ensures that the EC model dynamically evolves, reflecting both the insights from LLMs and the iterative learning process.

By leveraging the flexibility and multidimensional approach of EC theory, we can transcend the traditional limitations posed by ANNs. This progressive modeling approach not only aligns with the complex nature of financial ecosystems but also offers a robust platform for developing advanced financial planning tools that are both comprehensive and capable of evolving in tandem with the changing economic landscape.

5. LLM Cooperative Budgeting and Financial Planning for Households

Cooperative budgeting emerges as an imperative when multiple individuals, such as partners or roommates, decide to combine their financial resources and manage expenses collectively. By jointly analyzing incomes and costs, households can better allocate resources, set shared financial goals, and plan for both short-term needs and long-term objectives.

The provided Python script, `get_cooperative_financial_recommendation`, seeks to tap into the potential of LLMs to generate budgeting advice tailored for households that are pooling their resources.

Key Script Features:

- Caching Mechanism: Repetitive queries are bypassed by maintaining a `recommendation_cache`. If data from a household match a prior request, the cached recommendation is returned, thereby conserving computational effort.
- Retry with Exponential Backoff: The script is robust against potential API timeouts or other exceptions. By employing a retry mechanism that increases the waiting time after each attempt, it provides resilience against transient issues.
- Tailored Recommendations: The response from the LLM is parsed to yield a succinct and actionable cooperative financial recommendation for the household.

Sample Use Case

Consider a household comprising two members:

- Member 1 with an income of USD 5000, paying rent of USD 800 and spending USD 300 on groceries.
- Member 2 earning USD 4000, having a rent of USD 700 and grocery expenses amounting to USD 350.

Upon feeding these combined data to the LLM, a recommendation could emerge as:

"Given the combined income and individual expenses, the household has a remaining amount of $7850 after rent and groceries. It is advised to allocate $1500 for joint savings, $1000 for utilities and shared amenities, $500 towards entertainment and leisure, and set aside $1000 as an emergency fund. The residual amount can be directed towards individual savings or other shared financial goals."

Benefits of Collaborative Budgeting:

1. Shared Responsibility: Distributing financial responsibilities can alleviate individual burdens and promote mutual accountability.
2. Strategic Decision Making: With pooled resources, households can make more informed choices about investments, large purchases, and savings.
3. Achievement of Common Goals: Be it buying a new home, planning a vacation, or preparing for family expansions, cooperative budgeting can expedite the realization of shared aspirations.

Enhancements for the Future:
The script, while effective, can be augmented further by:

- Incorporating feedback mechanisms where household members can adjust recommendations based on changing financial scenarios.
- Integrating data from financial institutions in real-time for more current and accurate recommendations.
- Introducing functionalities to handle more complex financial situations, such as investments, loans, and mortgages.

In essence, cooperative budgeting facilitated by AI-driven models can revolutionize the way households manage their finances, promoting a culture of transparency, mutual trust, and shared financial growth.

6. Cooperative Budgeting within the EC Framework

Cooperative budgeting for households represents a tangible instantiation of the extended coevolutionary (EC) framework's principles first introduced in [27]. By analyzing this in the context of the EC theory, we illuminate the synergistic confluence of game-theoretic models, LLMs, and adaptive learning in practical MAS scenarios.

Let us consider a household H consisting of n members, where each member o is an agent. Each agent has a financial profile characterized by its income I_o and expenses E_o. The collective budget of the household can be expressed as:

$$B_H = \sum_{o=1}^{n} I_o - E_o. \tag{26}$$

The goal is to cooperatively allocate a common budget B_H across members to optimize some household-wide utility function U_H. This utility can be a measure of financial stability, savings, or any other metric significant to the household.

Incorporating the adaptive learning mechanism from the EC framework, each agent o iteratively adjusts its expenses E_o based on a derived utility U_o that takes into account the household's collective goal and individual preferences. The adjustment can be mathematically represented analogous to the previous equation as:

$$E_o(t+1) = E_o(t) + \alpha \nabla U_o(E_o(t), E_{-o}(t)). \tag{27}$$

Here, α remains the learning rate, and ∇U_o signifies the gradient of utility with respect to agent o's expenses, considering other agents' expenses as E_{-o}.

The strategy of agent o, specifically its expense E_o, can be influenced by recommendations from an LLM. Representing the LLM's recommendation for agent o at time t as $R_{o,t}$, and similar to the equation presented before, the new expense can be modeled as:

$$E_o(t+1) = (1-\beta)(E_o(t) + \alpha \nabla U_o(E_o(t), E_{-o}(t))) + \beta R_{o,t}. \tag{28}$$

Here, β designates the influence rate of the LLM on the agent's expense strategy.

Given that LLM recommendations, like budget suggestions, come with inherent uncertainties, a confidence measure $C_{o,t}$ associated with each recommendation $R_{o,t}$ should be considered. This is analogous to the previous equation, and the updated expense can be modeled as:

$$E_o(t+1) = (1 - \beta C_{o,t})(E_o(t) + \alpha \nabla U_o(E_o(t), E_{-o}(t))) + \beta C_{o,t} R_{o,t}. \tag{29}$$

Each household member may have distinct financial priorities and constraints, echoing the EC theory's emphasis on heterogeneous agents. The utility for each agent U_o would thus encapsulate individual preferences, like risk aversion or future financial goals.

The cooperative budgeting scenario for households aptly demonstrates the applicability of the EC theory in real-world contexts. This intersection of theoretical constructs with tangible applications showcases the robustness and versatility of the EC framework. The use of mathematical representations deepens our understanding, offering a quantitative lens to perceive the dynamics of the household as a multi-agent system influenced by adaptive learning and LLMs.

Linear Optimization Model for Financial Planning with LLM Constraints

Another instance of the same financial planning scenario is the following, where we propose to use the LLM recommendation as bounds of the optimization model. We construct a linear optimization model to distribute the monthly income I across various expense categories while maximizing potential savings. The model is formulated as follows:

Let I represent the monthly income of a household, and E_z represent the expenses in category z, where z can be rent, groceries, utility bills, or entertainment. Let S denote the savings for the month. The goal is to maximize S subject to the constraints imposed by the financial data and recommendations from the LLM.

The objective function and constraints are presented as follows:

$$\text{Maximize } S = I - \sum_z E_z \tag{30}$$

$$E_z \geq 0 \, \forall z, \qquad \text{(Non-negativity of expenses)} \tag{31}$$

$$S \geq 0, \qquad \text{(Non-negativity of savings)} \tag{32}$$

$$\sum_z E_z \leq I, \qquad \text{(Total expenses do not exceed income)} \tag{33}$$

Additional constraints are derived from the LLM recommendations, which suggest percentages of income to be allocated toward each category. Let p_z be the percentage of income recommended for category z, and E_z^{rec} be the recommended expenses calculated as $p_z \cdot I$. We can introduce bounds based on these recommendations to guide the optimization:

$$E_z \geq E_z^{\min}, \qquad \text{(Minimum spending)} \tag{34}$$

$$E_z \leq E_z^{\text{rec}}, \qquad \text{(LLM-recommended spending)} \tag{35}$$

where $E_z^{\min}$ is a lower bound to ensure that each category meets essential spending requirements. The upper bound E_z^{rec} is influenced by the LLM's advice and represents an ideal spending target as suggested by the model.

7. Simulation Framework: LLM-Informed Decisions

To comprehend the LLM's effectiveness in financial advice, it is essential to assess its recommendations. Let us delve into its suggestions for individual and cooperative financial management:

7.1. User Financial Recommendation

OpenAI model: gpt4-0613

```
response = openai.ChatCompletion.create( model="gpt-4-0613", messages=[ {"role": "system", "content": "You are an expert financial planner with extensive knowledge in household finance and budgeting. Provide a detailed recommendation on how to allocate funds based on the given data."}, {"role": "user", "content": f"Here's my financial data for the month: {income_data_string}. Can you provide a recommendation on how I should spend or save the remaining funds?"} ] )
```

The instruction given to the model focused on its expertise as a financial planner. This query is indicative of how the LLM would process individual financial data and provide a comprehensive recommendation.

LLM User Financial Recommendation

Of course, based on your provided income and expenditure, your total spending adds up to $1900 out of $5000. This leaves you with a remaining balance of $3100. Let's break down and allocate your remaining balance as follows:
1. **Emergency Fund**: If you don't already have an emergency fund, aim to build one. This should consist of 3–6 months' worth of living expenses to cover unexpected costs, such as a medical emergency or car repair. If you don't have such a fund pot, then it's good to allocate around 20% ($620) of the remaining money towards it.
2. **Retirement Savings**: If you're not currently contributing to a retirement fund, it's a good idea to start. A general rule is to allocate 15–20% of your income towards your retirement. In this case from the remaining sum, consider allocating 20% ($620).
3. **Debt Payments**: If you have any outstanding debt like student loans, credit card debt, etc, it's crucial to allocate a portion of your income towards paying these off. For now, from this remaining balance you could allocate 15% ($465).
4. **Personal Savings/Investments**: It's always a good idea to invest in your future, whether it means buying a house, starting your own business, or taking a vacation. You could aim to allocate 20% ($620) for this purpose.
5. **Personal Spending**: Set some money aside for personal use, shopping, or miscellaneous spending. Around 10% ($310) could be allocated for this purpose.
6. **Health and Fitness**: Consider this as an investment in your health. Allocate about 5% ($155).
7. **Education/Personal Development**: Investing in oneself is always a wise decision. 5% of the remaining money which would be around $155 could be dedicated for this purpose.
8. **Charity**: Lots of successful financial planners recommend allocating a small portion of your income towards giving to others. Around 5% ($155) should be again a good start.
Here's how the final allocation looks like:

- Emergency Fund: $620
- Retirement Savings: $620
- Debt payments: $465
- Personal Savings/Investments: $620
- Personal Spending: $310
- Health and Fitness: $155
- Education/Personal Development: $155
- Charity: $155

Analysis: The recommendation begins by acknowledging the user's financial state, ensuring transparency in the LLM's calculations. The model then meticulously divides the

leftover funds into eight categories. These categories represent the fundamental areas that most individuals need to consider in their financial planning. The proportional allocation seems balanced and encourages both saving and investing while also emphasizing personal well-being and community giving.

This framework aligns well with traditional financial planning advice, indicating the model's efficacy in this domain. Moreover, the clarity and structure provided by the model can greatly assist users in understanding their financial standing and planning their future expenditures.

7.2. Cooperative Financial Recommendation

OpenAI model: gpt4-0613

```
response = openai.ChatCompletion.create( model="gpt-4-0613", messages=[ {"role": "system", "content": "You are an expert financial planner with expertise in cooperative budgeting for households. Provide a comprehensive recommendation based on the combined financial data."}, {"role": "user", "content": f"Here's our combined household financial data: {household_data_string}. Can you recommend how we should collaboratively budget our expenses?"} ] )
```

In this inquiry, the model was tasked with providing a comprehensive recommendation based on combined financial data, embodying the complexities of cooperative budgeting for households.

Analysis: The model adopts a detailed approach to cooperative financial planning, starting with a clear breakdown of combined income and essential expenses like housing and groceries. The percentages used to allocate various expenses, such as housing taking up 25–30% of the income, align with widely accepted financial guidelines.

A particularly noteworthy aspect of this recommendation is the inclusion of an emergency fund, highlighting the LLM's capacity to foresee potential financial pitfalls and its alignment with prudent financial practices.

The cooperative recommendation, much like the individual one, is comprehensive, detailed, and aligns with standard financial planning guidelines. This reinforces the model's reliability in producing accurate and actionable financial advice.

LLM Cooperative Financial Recommendation

Absolutely. Here's a comprehensive budgeting plan based on the combined financial data:
1. **Income Analysis:** Your combined total monthly income is $5000 + $4000 = $9000.
2. **Housing Expenses:** Your combined rent is $800 + $700 = $1500, which is 16.67% of your income. Considering that housing should generally be around 25–30% of your income, you're in a good position here.
3. **Grocery Expenses:** Your combined grocery cost is $300 + $350 = $650 per month. This cost is approximately 7.2% of your total income which is below the usual 10–15% range. You are so far managing this expense effectively but there may be room for improved nutrition or desired treats within this budget.
4. **Savings:** A good rule of thumb would be to save at least 20% of your income. In your case, that would be $1800 per month.
5. **Other Expenses:** After accounting for rent, groceries and savings, you have $4050 left each month ($9000 − $1500 − $650 − $1800). You should use this remaining money to cover other expenses. It would be advisable to distribute it as follows:

- **Utilities: 5–10% of income:** This would be about $450–$900.
- **Healthcare: 5–10% of income:** This should be approximately $450–$900.
- **Transportation: 10–15% of income:** This should ideally be between $900–$1350.
- **Recreational & Personal Expenses: 5–10% of income:** This is typically $450–$900.
- **Retirement Contribution: At least 10% of income:** This would be $900 in your case.

This distribution may vary based on your personal needs, lifestyle, debt repayments, and obligations. It would be wise to keep track of your expenses to ensure you're in-line with your set budget.
6. **Emergency Fund:** It is also important to ensure that you have an emergency fund in place. Ideally, this fund should cover 3–6 months' worth of living expenses, and given your expense rate, this should be somewhere between $10,800–$21,600. If you do not have an adequate emergency fund in place, then portion of your savings should be allocated to this fund until the target is met.
End of the day, the goal is to ensure the income is distributed in a way that covers all the needs, while also ensuring there's financial growth and stability. Regularly revisit your budget and adjust accordingly based on changes in income, expenses, and financial goals.

7.3. The Significance of Long-Term Models in Financial Planning

Financial planning [28,29], at its core, is not just about budgeting for the next month or year: it is about laying a foundation for the entire life-cycle [30,31]. Decisions made today will inevitably impact financial security, flexibility, and wealth accumulation in the distant future. With this understanding, there arises an imperative need for long-term models in the domain of financial planning.

Long-term models bring foresight to financial decision making. By analyzing long-term implications of current financial data, these models offer recommendations that are not just about immediate benefit but also cater to future financial health. The importance of long-term models becomes even more pronounced in situations that involve uncertainties like changing job markets, economic fluctuations, or personal life events.

Here is a glimpse into the key benefits and functionalities of long-term models in the financial domain:

- Consumption Smoothing: Long-term models assist in balancing consumption patterns over time, ensuring that individuals do not overspend during prosperous periods or

face hardship during lean times. By forecasting future income streams and financial obligations, long-term models promote consistent living standards throughout an individual's life.

- Life-cycle Modeling: These models simulate various stages of an individual's life, from early career stages, potential family planning, and mid-life crises, to retirement. By doing so, they help in preparing for financial needs and challenges typical to each phase.
- Risk Management: LLMs highlight potential risks by analyzing economic trends, inflation rates, and personal financial data. They help to devise strategies to mitigate these risks, be it through investments, insurance, or savings.
- Investment Guidance: Based on long-term financial goals and market predictions, LLMs provide guidance on where and how much to invest. This includes suggestions on diversifying portfolios, adjusting asset allocations, and more.
- Future Projections: One of the standout features of LLMs is their ability to project financial futures based on current data and decisions. They offer a roadmap of where one's finances could be in 5, 10, or even 30 years from now.

7.4. Long-Term Cooperative Financial Recommendation

OpenAI model: gpt4-0613

```
response = openai.ChatCompletion.create( model="gpt-4-0613", messages=[ {"role": "system", "content": "You are an expert financial planner with expertise in life-cycle models and cooperative budgeting for households. Provide a comprehensive recommendation based on the combined financial data that considers consumption smoothing over the entire life-cycle."}, {"role": "user", "content": f"Here's our combined household financial data, along with some future projections and life events: {household_data_string}. Can you recommend how we should collaboratively budget our expenses?"} ] )
```

Here is an example usage with more comprehensive data taking into account long-term dependencies:

LLM Long-Term Cooperative Financial Recommendation

Using the life-cycle model involves setting your saving and spending habits with a long-term vision. Consumption smoothing is a key part of this, aiming to allocate resources in a way that maintains a consistent standard of living throughout your life. Here are my recommendations based on both your income and projected life events:
1. **Start an Emergency Fund**: Save a portion of your combined income consistently until you have about three to six months' worth of living expenses. Considering your incomes and basic monthly expenses, which total $2150, I'd suggest allocating at least $500/month for this initially.
2. **Create a Child Fund**: Since you're planning to have a child in 3 years, creating a dedicated child fund should be a priority. It will help cover everything from pregnancy-related medical costs to future schooling expenses. Let's allocate $300/month for this fund.
3. **Retirement Saving**: Member 2 is planning to retire in 20 years, so a substantial part of the income should go towards retirement savings. If not already started, I propose an immediate beginning with an initial tier of at least $600 a month.
4. **Income Adjustment**: The potential pay cut for Member 1 should be planned for in advance. To facilitate consumption smoothing, consider increasing savings or reducing non-essential expenses to accommodate the projected 10% pay cut.
5. **Housing Expenses**: Your total monthly rent is $1500. If you stay in the current rental situation for the long term, there could be an opportunity to buy a house, as long as it's financially reasonable.
6. **Daily Living and Unexpected Expenses**: Allocate the rest of the income for your daily living expenses and unexpected costs (car repairs, medical expenses, etc.).
In summary, this is a rough recommendation on which categories can the budget be divided:

- **Emergency Fund**: $500/month
- **Child Fund**: $300/month
- **Retirement Savings**: $600/month
- **Housing**: $1,500/month
- **Groceries**: $650/month
- **Daily Living and Unexpected Costs**: The balance of your incomes after removing the above expenses (approximately $2450).

This plan will require review and adjustment on a regular basis, at least yearly, to accommodate changes in income or new financial needs/goals. Remember, the key to the life-cycle model is flexibility and long-term planning.

As seen from our simulation framework, when LLMs like gpt-4-0613 are given a long-term perspective, they provide advice that aligns with the core tenets of financial planning. Whether it is individual or cooperative budgeting, the model's recommendations reflect an understanding of the intricate dance between current spending, future needs, and long-term financial growth.

Analysis: The long-term cooperative financial recommendation exemplifies the model's capacity to engage with complex financial scenarios that span across an individual's lifetime. Notably:

- Consumption Smoothing: The model emphasizes consumption smoothing by suggesting a consistent standard of living. Its advice on adjusting income to accommodate a projected pay cut is particularly salient, as it showcases the model's ability to suggest proactive measures against potential financial setbacks.
- Emergency and Specific Funds: The inclusion of both an emergency fund and a child fund showcases the model's comprehensive approach. It highlights an understanding

of immediate needs (like emergencies) and future events (like planning for a child), thus promoting financial preparedness.

- Retirement Planning: The model's advice to prioritize retirement savings for Member 2, who plans to retire in 20 years, underscores its alignment with a long-term vision. It demonstrates an acute awareness of the time-sensitive nature of retirement planning.

Overall, the model's long-term cooperative financial recommendation reflects a good understanding of the life-cycle model's principles. It meticulously integrates immediate financial needs with long-term goals, showcasing its potential to serve as a valuable tool in the domain of financial planning.

7.5. Simulation of Financial Planning with LLM Recommendations

Going beyond, we present a simulation that integrates financial recommendations from an LLM into the financial planning process of synthetic household datasets. Our approach involves generating synthetic financial profiles for ten households, encompassing monthly income and expenses across four categories: rent, groceries, utility bills, and entertainment. The monthly incomes are normally distributed around a mean of USD 5000 with a standard deviation of USD 1000. Expenses in each category are also generated following a normal distribution with category-specific means and variances.

Once the synthetic dataset is established, we employ an LLM to provide financial planning advice for each household. The LLM is prompted to analyze the financial data and recommend allocations as percentages of the household's monthly income. The advice is structured to include essential expenditures, savings, and discretionary spending, following sound financial planning principles.

The LLM's textual recommendations are parsed using regular expressions to extract the suggested allocation percentages for each financial category. These extracted values are then translated into monetary amounts based on the respective household's income and stored alongside the original data.

Subsequently, we apply a linear programming optimization model to each household's financial data, aiming to maximize savings while adhering to realistic constraints. This model takes into account minimum spending requirements for each expense category, ensuring that the optimized plan remains practical and sustainable.

Finally, we visualize the simulation results, comparing the original savings, LLM-recommended savings, and optimized savings with bounds. This comparison highlights the potential impact of AI-driven financial advice when combined with traditional optimization techniques.

The instance of a financial planning scenario in Figure 2 not only showcases the feasibility of utilizing AI in personal financial planning but also serves as a benchmark for evaluating the LLM's performance against traditional optimization models. The findings suggest that LLMs can effectively contribute to financial decision making, offering insights that align with recognized financial best practices.

The simulation involves several stages:

1. Generation of synthetic financial data for each household, with random distributions for income and expenses.
2. Retrieval of financial recommendations from an LLM, which analyzes the synthetic data and provides allocation percentages for various expense categories.
3. Parsing of the LLM output to translate percentage-based recommendations into monetary allocations.
4. Application of a linear optimization model to determine an optimized savings strategy for each household, independent of LLM recommendations, using fixed minimum spending bounds for each expense category.
5. Visual comparison of the households' original savings, the savings as recommended by the LLM, and the savings post-optimization, highlighting the potential of LLM guidance in enhancing financial decision making.

LLM Recommendation

LLM_output = Based on your income and expense data, here is how I will suggest allocating your funds:

- Rent: 24% of income
- Groceries: 8% of income
- Utility Bills: 4% of income
- Entertainment: 2% of income
- Emergency Fund: 20% of income
- Retirement Savings: 15% of income
- Non-retirement Savings: 10% of income
- Debts: 8.5% of income
- Personal Spending: 8.5% of income

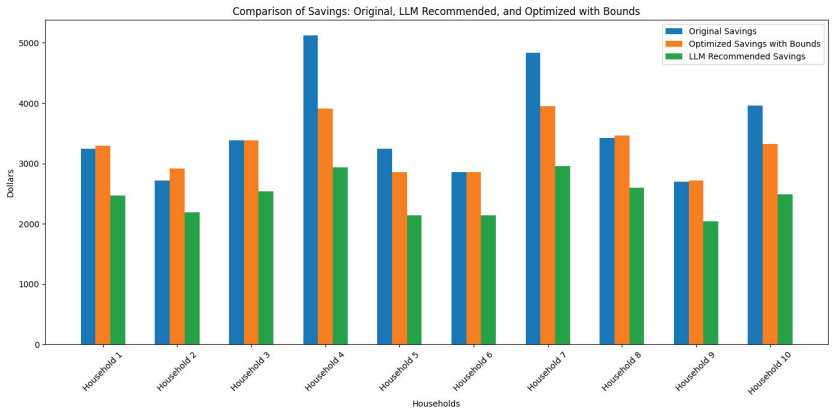


Figure 2. Comparative analysis of household savings across three different financial planning scenarios. The 'Original Savings' bars represent the savings calculated from the synthetic dataset without any optimization. The 'Optimized Savings with Bounds' bars illustrate the savings after applying a linear optimization model that maximizes savings while considering minimum essential expenditures. The 'LLM Recommended Savings' bars reflect the aggregate of savings categories as suggested by the LLM, including emergency fund, retirement savings, and non-retirement savings.

In the process of developing our LLM financial recommender system, we mapped the interactions between various components to ensure a seamless user experience and robust performance. Figure 3 illustrates the sequence of events from the moment a user inputs their financial data to the receipt of personalized recommendations and subsequent budget optimization. This diagram delineates the system's workflow, highlighting the role of the LLM in analyzing financial data and generating tailored advice. The sequence further demonstrates the integration of an optimization model that fine-tunes the initial recommendations to accommodate user-specific financial goals and constraints. The interaction between the user, LLM, and optimization model encapsulates the synergy of AI-driven insights with algorithmic precision, which is central to the innovative approach of our financial planning system.

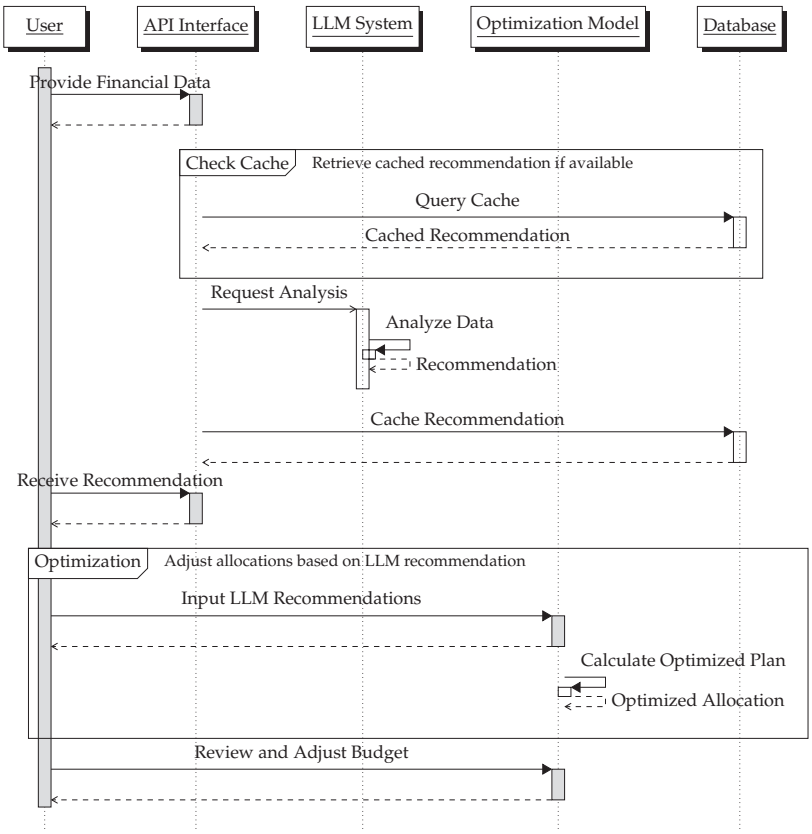


Figure 3. Sequence diagram of the LLM financial recommender system illustrating the interaction between the user, LLM system, API interface, optimization model, and database.

The qualitative analysis in Table 1 distinctly showcases the advantages of our proposed models over traditional financial planning methodologies. Key enhancements include the capacity for high personalization, where the LLM's inferential abilities tailor recommendations to specific user scenarios. Unlike traditional models, our approach dynamically adapts to market changes, leveraging the LLM's capacity to analyze current data trends. Scalability is another significant improvement: the AI-driven model can simultaneously cater to a vast array of users, a feat challenging for traditional methods. In terms of user accessibility, the proposed system simplifies complex financial concepts, making financial planning more approachable for the general public. The integration of real-time data ensures that the advice is always relevant and timely, a critical aspect often lacking in conventional models. Finally, the proposed methodology excels in long-term financial planning, with the ability to project and adapt to future financial scenarios, providing users with a comprehensive view of their financial trajectory.

Table 1. Comparative analysis of traditional and proposed LLM-integrated models.

Criteria	Traditional Models	Proposed Models with LLM Integration	Comments
Personalization	Low	High	The integration of LLMs allows for highly personalized financial planning based on individual or household-specific data.
Adaptability to Market Changes	Limited	Enhanced	LLMs can rapidly adjust recommendations in response to fluctuating market conditions.
Scalability	Moderate	High	The AI-driven approach can easily scale to handle a large number of users simultaneously.
User Accessibility	Variable	Improved	The use of LLMs makes financial planning more accessible to users without deep financial knowledge.
Real-time Data Integration	Limited	Enhanced	Proposed models can integrate and analyze real-time data for up-to-date advice.
Long-term Financial Planning	Moderate	Superior	LLMs can project long-term financial outcomes, aiding in more effective life-cycle financial planning.

8. Conclusions and Future Work

The transition from traditional optimization approaches to the use of the extended coevolutionary (EC) theory presents an innovative direction for the domain of financial planning. By analyzing cooperative budgeting within households, this paper illuminates the compatibility of game-theoretic models, adaptive learning, and the influential role of large language models (LLMs) in shaping financial recommendations.

The provided pipeline exemplifies practical applications of the EC theory [27], demonstrating a marriage between adaptive learning mechanisms and LLMs to generate financial advisories, as an instantiation of a new economic theory where human decisions can be greatly influenced by AI agents. Through iterative refinements influenced by LLM recommendations, the system iteratively approaches optimal financial allocations. The clear distinction between individualized financial recommendations and cooperative household budgeting reinforces the versatility of this approach, which could be seamlessly integrated into household members by the use of a smartphone, smartwatch, or personal computer.

Future work may involve:

- Expansion of Utility Functions: Future studies could delve deeper into constructing more comprehensive utility functions, encompassing factors like changing financial landscapes, inflation rates, and socio-economic considerations.
- Model Robustness: Analyzing the robustness of LLM recommendations in diverse financial scenarios, especially under market volatility, would be beneficial. This ensures the system's resilience and adaptability to external shocks.
- Integration with Real-time Data: By interfacing the proposed methodology with real-time financial databases, the recommendations can be rendered even more timely and relevant.
- Human-in-the-loop Mechanism: While LLMs are proficient, human expertise remains indispensable. Incorporating a human-in-the-loop mechanism, where financial experts refine or validate LLM recommendations, could elevate the system's reliability.
- Ethical Considerations: As with any AI-driven system, ethical considerations, especially in a domain as crucial as finance, are paramount. Future work must ensure that the recommendations are unbiased, transparent, and in the best interest of the users.

- Scalability to Larger Cooperative Entities: While households were the primary focus, scaling the model to address cooperative budgeting for larger entities, like community groups or small businesses, could be a natural progression.

In essence, the integration of the EC framework into financial planning is but the tip of the iceberg. The potential applications and refinements are vast, opening avenues for research and practical implementations that can significantly enhance the quality of financial advisories and decision-making processes.

Author Contributions: Author contributions: conceptualization, J.d.C. and I.d.Z.; funding acquisition, G.R., C.T.C., and I.d.Z.; investigation, I.d.Z. and J.d.C.; methodology, I.d.Z., J.d.C., and G.R.; software, J.d.C. and I.d.Z.; supervision, C.T.C. and G.R.; writing—original draft, J.d.C. and I.d.Z.; writing—review and editing, J.d.C., I.d.Z., C.T.C., and G.R. All authors have read and agreed to the published version of the manuscript.

Funding: We thank the following funding sources from GOETHE-University Frankfurt am Main; 'DePP—Dezentrale Plannung von Platoons im Straßengüterverkehr mit Hilfe einer KI auf Basis einzelner LKW', 'Center for Data Science & AI', and 'xAIBiology'. We acknowledge the support of R&D project PID2021-122580NB-I00, funded by MCIN/AEI/10.13039/501100011033 and ERDF.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare that they have no conflicts of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

Large Language Models	LLMs
Deep Learning	DL
Artificial Neural Networks	ANNs
Multi-Agent Systems	MASs
Artificial Intelligence	AI
Extended Coevolutionary	EC
Retrieval Augmented Generation	RAG

References

1. Campbell, J.Y. Household finance. *J. Financ.* **2006**, *61*, 1553–1604. [CrossRef]
2. Guiso, L.; Sodini, P. Household finance: An emerging field. In *Handbook of the Economics of Finance*; Elsevier: Amsterdam, The Netherlands, 2013; Volume 2; pp. 1397–1532.
3. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30, (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
4. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021.
5. Perry, V.G.; Martin, K.; Schnare, A. Algorithms for All: Can AI in the Mortgage Market Expand Access to Homeownership? *AI* **2023**, *4*, 888–903. [CrossRef]
6. Roccetti, M.; Delnevo, G.; Casini, L.; Mirri, S. An alternative approach to dimension reduction for pareto distributed data: A case study. *J. Big Data* **2021**, *8*, 39. [CrossRef] [PubMed]
7. Markowitz, H. Portfolio Selection. *J. Financ.* **1952**, *7*, 77–91 . . [CrossRef]
8. Ando, A.; Modigliani, F. The 'Life Cycle' Hypothesis of Saving: Aggregate Implications and Tests. *Am. Econ. Rev.* **1963**, *53*, 55–84. Available online: http://www.jstor.org/stable/1817129 (accessed on 8 November 2023).
9. Charnes, A.; Cooper, W.W. Management Models and Industrial Applications of Linear Programming. *Manag. Sci.* **1957**, *4*, 38–91. [CrossRef]

10. Merton, R.C. Lifetime Portfolio Selection under Uncertainty: The Continuous-Time Case. *Rev. Econ. Stat.* **1969**, *51*, 247–257. [CrossRef]
11. Thaler, R.H. Toward a Positive Theory of Consumer Choice. *J. Econ. Behav. Organ.* **1980**, *1*, 39–60. [CrossRef]
12. Kahneman, D. Maps of Bounded Rationality: Psychology for Behavioral Economics. *Am. Econ. Rev.* **2003**, *93*, 1449–1475. [CrossRef]
13. Dixon, M.; Klabjan, D.; Bang, J.H. Classification-based Financial Markets Prediction using Deep Neural Networks. *Algorithmic Financ.* **2017**, *6*, 67–77. [CrossRef]
14. Mhlanga, D. Industry 4.0 in Finance: The Impact of Artificial Intelligence (AI) on Digital Financial Inclusion. *Int. J. Financ. Stud.* **2020**, *8*, 45. [CrossRef]
15. Althnian, A. Design of a Rule-Based Personal Finance Management System Based on Financial Well-Being. *Int. J. Adv. Comput. Sci. Appl.* **2021**, *12*, 182–187. [CrossRef]
16. Nazareth, N.; Reddy, Y.Y.R. Financial Applications of Machine Learning: A Literature Review. *Expert Syst. Appl.* **2023**, *219*, 119640. [CrossRef]
17. Dowling, M.; Lucey, B. ChatGPT for (Finance) Research: The Bananarama Conjecture. *Financ. Res. Lett.* **2023**, *53*, 103662. [CrossRef]
18. Weber, P.; Carl, K.V.; Hinz, O. Applications of Explainable Artificial Intelligence in Finance—A Systematic Review of Finance, Information Systems, and Computer Science Literature. *Manag. Rev. Q.* **2023**, 1–41. . [CrossRef]
19. Kumar, S.; Lim, W. M.; Sivarajah, U.; Kaur, J. Artificial Intelligence and Blockchain Integration in Business: Trends from a Bibliometric-Content Analysis. *Inf. Syst. Front.* **2023**, *25*, 871–896. [CrossRef] [PubMed]
20. Pallathadka, H.; Ramirez-Asis, E.H.; Loli-Poma, T.P.; Kaliyaperumal, K.; Ventayen, R.J.M.; Naved, M. Applications of Artificial Intelligence in Business Management, E-commerce and Finance. *Mater. Today Proc.* **2023**, *80*, 2610–2613. [CrossRef]
21. Arsenyan, J.; Piepenbrink, A. Artificial Intelligence Research in Management: A Computational Literature Review. *IEEE Trans. Eng. Manag.* **2023** . . [CrossRef]
22. Xiao, J.J. Applying behavior theories to financial behavior. In *Handbook of Consumer Finance Research*; Springer: New York, NY, USA , 2008; pp. 69–81. ._5. [CrossRef]
23. Browning, M.; Lusardi, A. Household Saving: Micro Theories and Micro Facts. *J. Econ. Lit.* **1996**, *34*, 1797–1855 . Available online: http://www.jstor.org/stable/2729595 (accessed on 8 November 2023).
24. Branzei, R.; Dimitrov, D.; Tijs, S. *Models in Cooperative Game Theory*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008; Volume 556.
25. Baker, H.K.; Filbeck, G.; Ricciardi, V. (Eds.). *Financial Behavior: Players, Services, Products, and Markets, Financial Markets and Investments*; Oxford Academic: New York, NY, USA, 2017.
26. Arner, D.W.; Zetzsche, D.A.; Buckley, R.P.; Barberis, J.N. FinTech and RegTech: Enabling Innovation While Preserving Financial Stability. *Georget. J. Int. Aff.* **2017**, *18*, 47–58. Available online: http://www.jstor.org/stable/26395923 (accessed on 8 November 2023). [CrossRef]
27. de Zarzà, I.; de Curtò, J.; Roig, G.; Manzoni, P.; Calafate, C.T. Emergent Cooperation and Strategy Adaptation in Multi-Agent Systems: An Extended Coevolutionary Theory with LLMs. *Electronics* **2023**, *12*, 2722. . [CrossRef]
28. Alhenawi, Y.; Hassan, M.K.; Hasan, R. Evolution of research in finance over the last two decades—A topographical view. *Res. Int. Bus. Financ.* **2022**, *59*, 101550. [CrossRef]
29. Luigi, G.; Luana, Z. From patriarchy to partnership: Gender equality and household finance. *J. Financ. Econ.* **2023**, *147*, 573–595.
30. Chai, J.; Horneff, W.; Maurer, R.; Mitchell, O.S. Optimal Portfolio Choice over the Life Cycle with Flexible Work, Endogenous Retirement, and Lifetime Payouts. *Rev. Financ.* **2011**, *15*, 875–907. . [CrossRef]
31. Browning, M.; Crossley, T.F. The Life-Cycle Model of Consumption and Saving. *J. Econ. Perspect.* **2001**, *15*, 3–22. . [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Journal of
Theoretical and Applied
Electronic Commerce Research

MDPI

Article

Enhancing the Prediction of Stock Market Movement Using Neutrosophic-Logic-Based Sentiment Analysis

Bassant A. Abdelfattah [1], Saad M. Darwish [1,*] and Saleh M. Elkaffas [2]

[1] Department of Information Technology, Institute of Graduate Studies and Research, Alexandria University, 163 Horreya Avenue, El-Shatby, Alexandria 21526, Egypt; igsr.bassant.a.abdelfattah@alexu.edu.eg
[2] Department of Information Systems, Arab Academy for Science and Technology, Alexandria 1029, Egypt; saleh.mesbah@aast.edu
* Correspondence: saad.darwish@alexu.edu.eg

Abstract: Social media platforms have allowed many people to publicly express and disseminate their opinions. A topic of considerable interest among researchers is the impact of social media on predicting the stock market. Positive or negative feedback about a company or service can potentially impact its stock price. Nevertheless, the prediction of stock market movement using sentiment analysis (SA) encounters hurdles stemming from the imprecisions observed in SA techniques demonstrated in prior studies, which overlook the uncertainty inherent in the data and consequently directly undermine the credibility of stock market indicators. In this paper, we proposed a novel model to enhance the prediction of stock market movements using SA by improving the process of SA using neutrosophic logic (NL), which accurately classifies tweets by handling uncertain and indeterminate data. For the prediction model, we use the result of sentiment analysis and historical stock market data as input for a deep learning algorithm called long short-term memory (LSTM) to predict the stock movement after a specific number of days. The results of this study demonstrated a predictive accuracy that surpasses the accuracy rate of previous studies in predicting stock price fluctuations when using the same dataset.

Keywords: long short-term memory; neutrosophic logic; sentiment analysis

1. Introduction

Predicting the stock market is an intricate challenge for investors, financial analysts, and researchers owing to its pivotal role in the global economy and its profound influence on individuals' lives [1,2]. Consequently, researchers have made extensive endeavors to improve stock market prediction through the utilization of various theories and employing machine learning methods or statistical modeling relying on diverse data sources [3]. The majority of previous studies have primarily relied on utilizing historical stock market data from platforms like Yahoo Finance. However, this approach poses significant limitations and shortcomings. It often fails to incorporate crucial hidden factors such as economic indicators, political events, investor sentiment, and market psychology [4,5]. As a result, this approach has exhibited considerable inaccuracy when attempting to forecast stock prices [6]. Thus, it is crucial to consider a wider range of factors and integrate alternative data sources such as Facebook, Twitter, now known as X, and news to improve the performance of stock market predictions [2].

Recently, the pervasive adoption of social media has facilitated the widespread dissemination of user experiences and feedback. This emerging trend has prompted researchers to investigate the utilization of social media as a valuable source for analyzing human sentiment concerning stock prices and capturing public opinions regarding specific companies or organizations. The incorporation of social media data into predictive models offers promising prospects for enhancing the accuracy and efficacy of stock market predictions [7].

Several research studies have primarily concentrated on examining the influence of sentiment analysis (SA) on stock market forecasting. However, these studies often overlook the presence of uncertainty or indeterminate data, which can significantly impact the accuracy of SA results. Neglecting such uncertainties can consequently affect the reliability and effectiveness of stock market predictions [8]. The neutrosophic concept, representing a broader perspective of fuzzy logic [9], has been introduced to tackle uncertainty and indeterminacy. It revolves around three membership functions, namely truth, indeterminacy, and falsity. These functions play a crucial role in representing and analyzing ambiguous and uncertain information [10,11].

1.1. Problem Statement

The financial market is an extremely complex domain, with the primary objective of accurately predicting stock movements. This precision is of the utmost importance as it directly influences the confidence of investors in their critical decisions regarding buying, holding, or selling stocks amid the inherent risks in the market. Predicting stock price movements remains a challenge despite extensive research efforts in forecasting using historical data and social media data reflecting user opinions about stocks and companies. However, it often lacks challenges, especially regarding the uncertainty in social media data, particularly the use of ambiguous words and multiple meanings. The presence of uncertainty in the data results in an inaccurate classification of the tweets, consequently leading to an inaccurate predictive model.

1.2. Motivation

The ability to accurately predict stock market movements has enormous benefits for investors and financial institutions. It can guide the investment decision-making process regarding buying or selling shares in response to price fluctuations. This can, in addition to reducing risk, promote financial stability and market efficiency. The stock market exhibits considerable volatility as it is influenced by various factors, including the impact of opinions and perceptions expressed on social media by users. However, one of the challenges of using social media is that data are sometimes uncertain and ambiguous. This motivates our research, which aims to improve stock movement prediction results by improving sentiment analysis results to tackle the challenge of data uncertainty in social media. This research will cover in detail how to detect data uncertainty and ambiguous opinions and how to resolve them using NL.

1.3. Contribution

The objectives of this paper are to enhance the accuracy of stock market movement prediction by integrating SA of public opinion from Twitter, and historical stock market data. This enhancement will be achieved by improving the sentiment analysis results, specifically designed to handle the uncertain and ambiguous data gathered from Twitter, about various companies and people's viewpoints regarding them. To address this challenge, we employ neutrosophic logic (NL) [12]. Furthermore, the output of sentiment analysis, alongside historical stock market data collected from Yahoo Finance, is incorporated into a model known as long short-term memory (LSTM) with the utmost accuracy to forecast the stock market movement. LSTM was chosen due to its capability to retain information over extended periods and demonstrate efficiency as a predictive model [13]. The core contributions of this paper are as follows:

- Proposal of a sentiment classifier based on NL to handle the ambiguity and uncertainty present in the data collected from Twitter, particularly concerning individuals' viewpoints regarding companies and stocks.
- Proposal of an enhanced prediction model that integrates the outcomes of sentiment classification and historical stock market data through the utilization of the LSTM model.
- The proposed model utilizes a benchmark dataset known as StockNet [14] to evaluate the efficiency of our proposed model compared to the other models.

The results of this study yielded a predictive accuracy of 78.48% in anticipating fluctuations in stock prices when utilizing the StockNet dataset [14]. Furthermore, this achievement surpassed the accuracy rate of prior studies that employed the same dataset.

The rest of this paper is structured as follows: Section 2 describes the related works of stock market prediction and NL. Section 3 focuses on the proposed model in detail. Section 4 illustrates the experimental results. The conclusion and future work are reported in Section 5.

2. Related Work

Predicting stock price trends has consistently captivated the interest of researchers, prompting investigations using various methodologies. Past endeavors in this field can be classified into two approaches: the analytical approach and the sentiment analysis approach. Within the analytical approach, the focus is on gathering historical stock data from various financial sources like Yahoo Finance and Google Finance. Conversely, the sentiment analysis approach is centered around users' reviews related to a specific stock or company, which are shared on diverse online platforms such as Facebook and Twitter [15]. This section provides a comprehensive literature review that primarily focuses on previous studies adopting the sentiment analysis approach. Moreover, we examine the influence of different techniques employed to enhance the accuracy of stock market forecasting and elucidate the correlation between stock market prices and social media or financial news [16]. In 2016, Pagolu et al. [17] tried to determine a correlation between the stock prices of Microsoft and public sentiments in tweets about that company to forecast the stock price for the next day. Two different methods, namely Word2vec and N-gram, were employed to calculate the polarity of each tweet. The accuracy of the polarity classification was approximately 70%, and the correlation between stock prices and sentiment was approximately 71.82%. However, it is worth noting that their study was limited by a small dataset comprising only 3216 tweets used for training the model.

In 2018, Xu et al. [14] suggested a deep generative approach, known as StockNet, for predicting stock market movement employing Twitter data and historical stock prices. This neural network architecture demonstrated a superior performance compared to prior works and incorporated recurrent, continuous latent variables to enhance the handling of randomness. However, the achieved accuracy was limited to approximately 58.23% due to the dataset's limited size. Another significant study in this direction was performed in 2019, where Xu et al. [18] suggested an attention-based LSTM approach that integrated tweets collected from StockTwits, historical data collected from Yahoo Finance, and technical indicators such as Average Directional Movement Index, Simple Moving Average, and Exponential Moving Average to enhance its effectiveness. The study primarily focused on comparing the performance of the attention-based LSTM approach with that of a traditional LSTM. The results of the model outperformed the traditional LSTM, achieving an accuracy rate of approximately 64%. However, it is important to note that this accuracy rate is still deemed insufficient due to the dataset's limited size, and more stocks and technical indicators need to be collected.

In 2020, Maqsood et al. [19] applied three machine learning algorithms, namely linear regression, support vector regression, and deep learning, to forecast stock market trends for the four primary stocks from the US, Turkey, Pakistan, and Hong Kong and evaluate the efficacy of each model. Each model incorporated Twitter data related to eight significant events to augment the accuracy of stock market predictions. The authors concluded that not all events have a direct influence on stock market prediction. Nevertheless, noteworthy local events possess the potential to impact the effectiveness of prediction algorithms. One of the disadvantages is that it was determined that the sentiment analysis technique (i.e., SentiWordNet) employed in their study was overly simplistic to establish comprehensive conclusions regarding this statement. In the same year, Gupta et al. [20] demonstrated the correlation between daily sentiments extracted from StockTwits and the corresponding daily movement of stock prices. Sentiment analysis was conducted using machine learning

methods (SVM, naïve Bayes, and logistic regression) and feature extraction techniques (bigram, bag of words, trigram, LSA, and TF-IDF). The fusion of TF-IDF and logistic regression achieved the highest accuracy, ranging from 75% to 85%. However, one study drawback was that the authors focused only on positive and negative sentiments while disregarding neutral sentiments.

Sawhney et al. [21] suggested a model, known as Multipronged Attention Network for Stock Forecasting (MAN-SF), which concurrently assimilates information from Twitter data, historical stock prices, and inter-stock relations for predicting stock market movement. The model leverages a graph neural network to discern relationships among different stocks, allowing it to acquire insights from correlations and interdependencies that impact stock movements. The authors employed the StockNet dataset and performed a comparative analysis of their model against others utilizing the same dataset. MAN-SF outperforms the strongest baselines, StockNet and Adversarial LSTM, demonstrating its superior performance in stock prediction with an accuracy of approximately 60.8%. Nevertheless, it is crucial to highlight that the accuracy rate was considered inadequate because of the dataset's limitations.

In 2021, Heiden et al. [13] tried to predict the stock price and investigated the impact of the news on it. They incorporated news sentiment as a feature in an LSTM prediction model, along with historical data. The authors obtained the news from the New York Times and employed a Valence Aware Dictionary and Sentiment Reasoner (VADER) for SA. The results indicated that including news sentiments improves the model's performance. Furthermore, the model exhibits promising possibilities to predict stock prices for approximately 60 days into the future. One of the disadvantages of the study was the utilization of a limited dataset, potentially leading to less reliable and accurate results.

Ho et al. [22] proposed a novel multi-channel collaborative network architecture for stock trend prediction. This architecture integrates social media sentiment features extracted from Twitter and candlestick chart features derived from the stock's historical time series data to capture temporal patterns and price dynamics. The network employs two branches, each employing specific deep-learning techniques. A one-dimensional convolutional neural network is utilized for sentiment classification on the extracted social media features, while two-dimensional convolutional neural networks perform image classification on the transformed candlestick chart data. The experimental results indicate a superior performance when compared to single-network models that rely solely on either candlestick charts or sentiment data. Notably, the proposed model achieved a prediction accuracy of 75.38% for Apple stock.

Another study in this area was achieved in 2022, where Fazlija et al. [23] demonstrated the use of financial market sentiment data extracted from news articles to predict fluctuations in the Standard & Poor's 500 stock market returns using sentiment values. The authors discovered that employing the Bidirectional Encoder Representations from Transformers (BERT) model yielded the highest success rate in sentiment classification. Moreover, they devised a random forest classifier technique for forecasting future price movements of the stock market index. The findings of the study demonstrated the significance of incorporating sentiment scores derived from news articles in forecasting the movement of stock prices. A notable limitation of the study is the omission of news data from all companies included in the stock market index.

Cristescu et al. [24] investigated the potential of SA to enhance the accuracy of the prediction of the stock market price using regression models. VADER was employed for SA based on news articles, and three types of regression models were implemented: cubic, quadratic, and linear regressions. The findings revealed that incorporating SA significantly improved the performance of the nonlinear regression model, evidenced by a superior fit compared to the linear model. Notably, the R-squared value was 0.005 for cubic regression and 0.001 for linear regression, indicating the superior performance of the cubic regression model.

Srijiranon et al. [25] tried to improve the prediction of the stock market by introducing a hybrid model that integrates three techniques: Principal Component Analysis, Empirical Mode Decomposition (EMD), and LSTM employing the incorporation of historical stock market data and the news. They utilized the Financial Bidirectional Encoder Representations from Transformers (FinBERT) for the SA process. The results indicated that the hybrid model demonstrated a superior performance, and the incorporation of news sentiment analysis enhanced the LSTM model's predictive capabilities.

Koukaras et al. [26] emphasized the significance of SA in the enhancement of stock market prediction by employing StockTwits and Twitter data. The authors integrated SA with machine learning and utilized seven machine learning algorithms, including Logistic Regression, Support Vector Machine, Multilayer Perceptron, k-nearest neighbors, naïve Bayes, Decision Tree, and Random Forest. They performed VADER and TextBlob for SA. The findings revealed that optimal results were achieved when employing VADER incorporating the Support Vector Machine with an F-score of 76.3%.

In 2023, Costola et al. [27] examined the correlation between the stock market and news regarding COVID-19 obtained from the New York Times, Reuters, and MarketWatch news platforms. To conduct SA, the authors employed a BERT model adapted for the financial market domain. The findings of this study revealed a positive relationship between the sentiment score and market returns. However, one limitation of the study is its exclusive focus on three news platforms, neglecting the inclusion of social networks as potential data sources. Nevertheless, none of the aforementioned approaches adequately capture the uncertainties and contradictions present in sentiment data. To tackle the issue of uncertainties in SA, the researchers attempted to leverage NL, aiming to enhance the efficiency of sentiment classification [28,29].

Kandasamy et al. [30] emphasized the presence of indeterminacy in the tweets based on the concept of neutrosophy. A dataset containing tweets related to the #MeToo movement was represented with positive, indeterminate, and negative memberships, forming a Set of Neutrosophic Values (SVNS) through the use of the VADER tool. Subsequently, the tweets were clustered and classified into positive, indeterminate, and negative classes. The K-means algorithm was employed to cluster the tuples into three major clusters, with the largest cluster representing the indeterminate tweets. To enhance the accuracy of predicting indeterminate polarity, the data were classified into eight distinct classes. The authors used training data to develop classifiers based on the Support Vector Machine (SVM), and k-nearest neighbor (k-NN) techniques. The study's findings revealed that the k-NN demonstrated superior performance compared to SVM in effectively classifying the SVNS values.

Kandasamy et al. [10] introduced a multi-refined neutrosophic set (MRNS), which refined the polarity result into seven classes (i.e., positive, indeterminate, negative, strong positive, indeterminate positive, indeterminate negative, and strong negative). The authors conducted a comparison between MRNS and two other approaches: one utilizing a single-valued neutrosophic set, and the other employing a triple-refined indeterminate neutrosophic set. The results demonstrated that MRNS outperformed the other approaches, providing a superior and more accurate result and effectively handling the inherent indeterminacy present in the data.

Reem et al. [12] introduced an opinion-mining model tailored for social media to tackle the challenges associated with ambiguous opinion classification. The proposed model incorporates social network analysis utilizing the UCINET tool, which is used for analyzing social network data (the University of California at Irvine Network), neural networks to assess the influence levels of users, and a classifier to combine their influence with the polarity of their texts. The authors evaluate the model using three classifiers (i.e., type-1 fuzzy logic, type-2 fuzzy logic, and NL) to handle the uncertainty in the data. The results demonstrated that NL surpasses the other classifiers in accuracy when dealing with data uncertainty. This highlights the effectiveness of NL in enhancing the performance of opinion mining in the realm of social media.

The Need to Extend the Related Work

Despite the existence of numerous studies on stock market prediction using SA, all these studies have demonstrated the influence of news and social media on the stock market. However, these studies primarily relied on traditional SA techniques and did not prioritize improving the SA results to address the challenges of data uncertainty.

In our work, we propose an enhanced model to predict the stock market movement using SA by proposing an advanced SA approach to deal with the ambiguity and uncertainty in the data by integrating a lexicon-based approach (i.e., VADER) with the NL technique. The key strength of NL is the ability to handle uncertainty by employing three membership functions, truth, indeterminacy, and falsity, to capture ambiguity and achieve accurate results closer to reality. To the best of our knowledge, this is the first time the NL technique is employed in the SA process to forecast the stock market movement.

3. Proposed Model

This model aims to enhance the prediction of stock market movements through the fusion of SA scores and historical stock data. This is achieved by improving the result of SA using NL, which can deal with indeterminacy in the data. The use of NL addresses the problem of data ambiguity when the tweet has the potential for two classifications. As a result, it improves the classification of tweets. The proposed model is discussed in the following steps:

- Proposal of an NL-based model integrated with a lexicon-based approach for calculating the sentiment classification of tweets to deal with uncertainty and ambiguity in the data collected from Twitter related to users' feedback and opinions about specific stocks and companies.
- Application of a deep learning algorithm LSTM to predict stock market price by using the result of sentiment score and the real stock price data as features in an LSTM model.
- Finally, the proposed model uses the integration of the SA score with the historical stock market. Moreover, it holds the distinction of being the first model that uses NL in the SA process to predict stock market movement.

Figure 1 presents the proposed model. The proposed model can be summarized in the following phases:

3.1. Data Collection

We have utilized a benchmark dataset called StockNet inspired by [14,21] which encompasses tweets and historical stock market data related to the stocks of 88 different companies over the period from 1 January 2014 to 1 January 2016. The tweets were obtained from Twitter based on NASDAQ ticker symbols (e.g., $APPL for Apple), while the historical data were collected from Yahoo Finance for the same period.

3.2. Data Pre-Processing

The purpose of this phase is to eliminate noise and irrelevant data from tweets to improve the effectiveness of the SA process. During this phase, several cleaning steps are performed [31]. Firstly, tweets are converted to lowercase. Subsequently, stopwords, punctuation, numbers, re-tweets, links, HTML tags, and special characters such as @ symbols, mentions, and hashtags are removed, ensuring a refined dataset for further analysis.

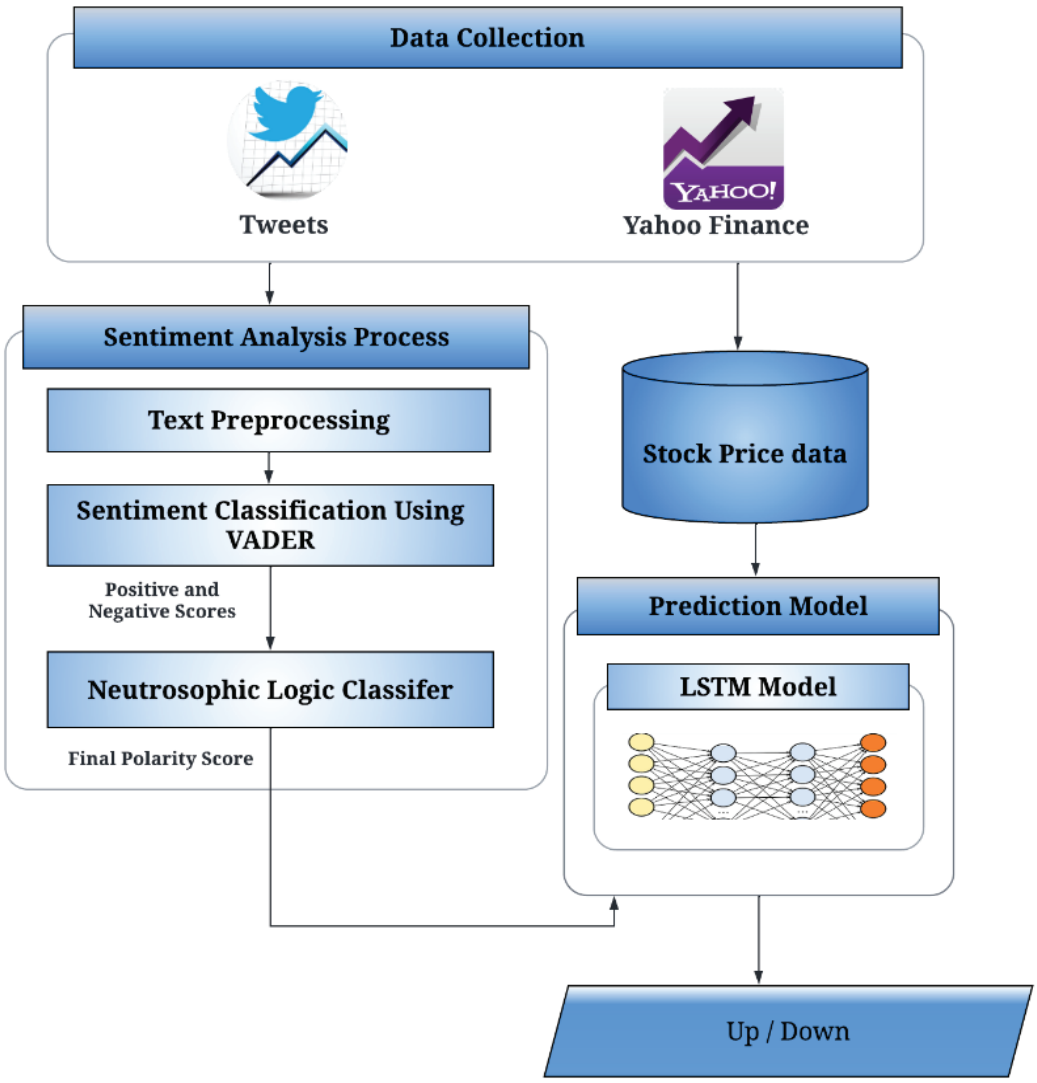


Figure 1. The proposed model.

3.3. Sentiment Analysis Based on Neutrosophic Logic

In this phase, we conduct sentiment classification of the tweets. and address the challenges of uncertainty and ambiguity in the social media data. The presence of uncertainty in the data poses a significant issue in sentiment analysis (SA), potentially resulting in data misclassification. Previous studies showed that both type 1 and type 2 fuzzy logic, as well as NL, can effectively address the uncertainty inherent in the data. Fuzzy logic, whether type 1 or type 2, introduces an overlapping region between classes, leading to ambiguity in the classification. Conversely, NL processing is adept at handling such ambiguity, thereby enhancing the efficiency of the classification process, as illustrated in Figure 2 [12,29,32,33]. Our aim is to enhance the efficiency of classification by using NL with a sentiment lexicon. The purpose of this phase is achieved through the following steps:

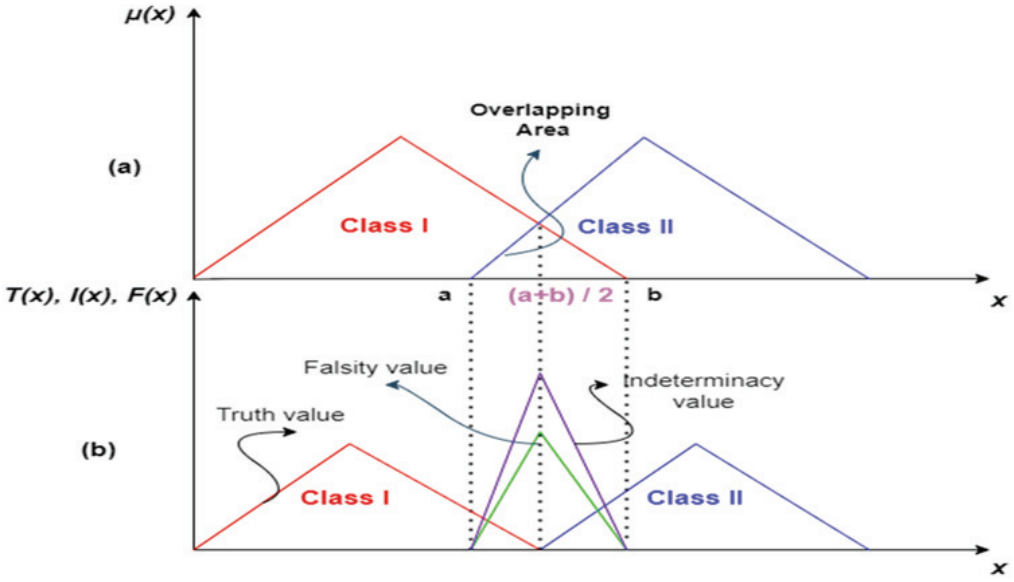


Figure 2. Comparison between fuzzy logic and NL: (**a**) Fuzzy logic membership function (**b**) NL truth, indeterminacy, and falsity membership functions.

3.3.1. Use of Lexicon-Based Approach

This process is concerned with computing the positive and negative scores of each tweet after the preprocessing phase using a lexicon-based sentiment technique (i.e., VADER) based on [34]. VADER is a Natural Language Toolkit (NLTK) Python package that generates four outputs: the scores of positive, negative, neutral, and compound (i.e., the total score of positive, neutral, and negative normalized between -1 and $+1$) for each text. VADER was selected due to its demonstrated superior performance when compared to other lexicons such as SentiWordNet, and TextBlob [25,33]. Additionally, it can provide positive and negative sentiment scores, which we incorporated into our NL classifier [34]. After applying VADER, only the positive and negative scores are fed into the NL classifier for further analysis.

3.3.2. Neutrosophic-Logic-Rule-Based Classification

The most important part of our work is this phase as it can solve the issue of ambiguity and uncertainty in the data and generate precise results that closely resemble how people interpret the texts. NL is defined by three membership functions (MF): truth membership function, indeterminacy membership function, and falsity membership function [12,32]. The output variables are represented as three components (truth, indeterminacy, and falsity), ensuring that there is no overlap between any two MFs [12]. The block diagram representation of an NL classification system is illustrated in Figure 3. The NL process is achieved in three steps, as follows:

(1) Neutrosophication

In this phase, the crisp inputs are transformed into neutrosophic sets using three triangular MFs: truth, indeterminate, and falsity. The inputs derived from the previous phase after applying VADER to the tweets include the positive and the negative score (PS and NS, respectively). The two inputs are assigned a value between 0 and 1, each with the levels low (L), moderate (M), and high (H) for the truth component according to Equation (1). Meanwhile, for the indeterminacy according to Equation (2) and falsity components according to Equation (3), the levels are low–moderate (L-M) and moderate–high (M-H) [12,33,35]. The output variable was assigned a value ranging from 0 to 1 with three classes (positive, neutral, and negative) inspired by the work introduced in [34]. Figure 4 illustrates the design of the NL truth MFs of the two inputs, PS and NS, whereas Figure 5 shows the indeterminate MFs, and Figure 6 illustrates the falsity MFs.

$$T(x) = \begin{cases} \alpha_{(\frac{x-a_1}{a_2-a_1})} & (a_1 \le x \le a_2) \\ \alpha & (x=a_2) \\ \alpha_{(\frac{a_3-x}{a_3-a_2})} & (a_2 \le x \le a_3) \\ 0 & otherwise \end{cases} \tag{1}$$

$$I(x) = \begin{cases} \frac{(a_2-x+\theta(x-a_1))}{a_2-a_1} & (a_1 \le x \le a_2) \\ \theta & (x = a_2) \\ \frac{(x-a_2+\theta(a_3-x))}{a_3-a_2} & (a_2 \le x \le a_3) \\ 1 & otherwise \end{cases} \tag{2}$$

$$F(x) = \begin{cases} \frac{(a_2-x+\beta(x-a_1))}{a_2-a_1} & (a_1 \le x \le a_2) \\ \beta & (x = a_2) \\ \frac{(x-a_2+\beta(a_3-x))}{a_3-a_2} & (a_2 \le x \le a_3) \\ 1 & otherwise \end{cases} \tag{3}$$

where represents $\langle (a_1, a_2, a_3); \alpha, \theta, \beta \rangle$ which is a neutrosophic set, α represents the maximum truth membership degree, θ indicates the minimum indeterminacy membership degree, and β represents the minimum falsity membership degree. α, θ and $\beta \in [0, 1]$. Additionally, $a_1 \le a_2 \le a_3$. These assumptions are based on [12,33,35].

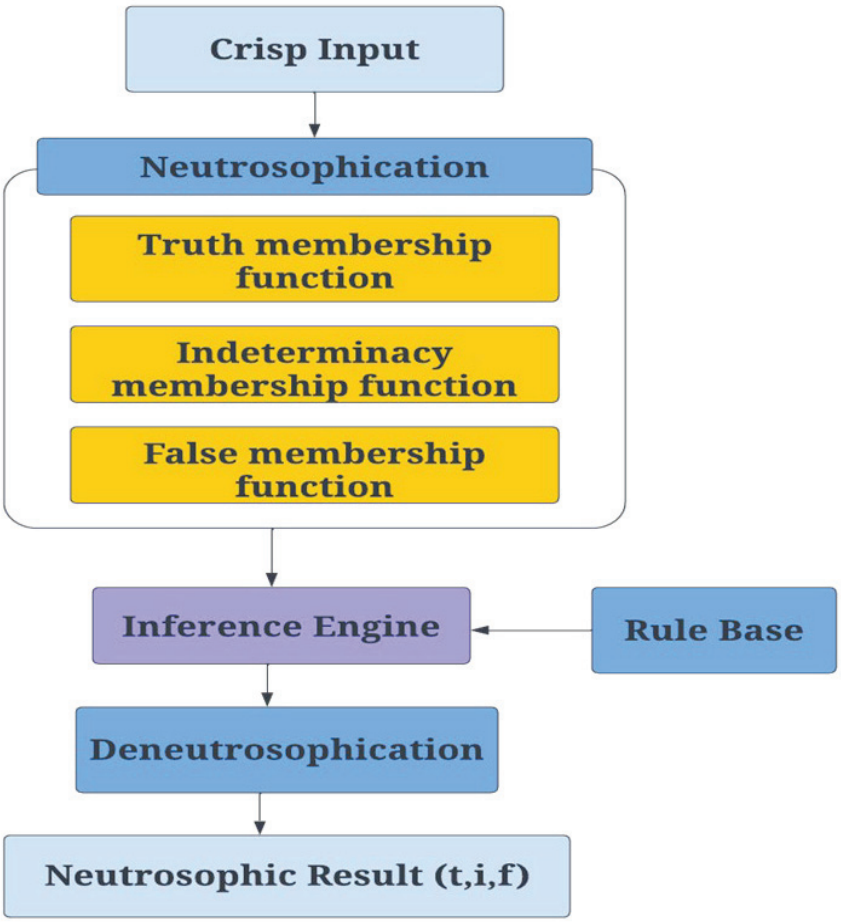


Figure 3. NL inference system.

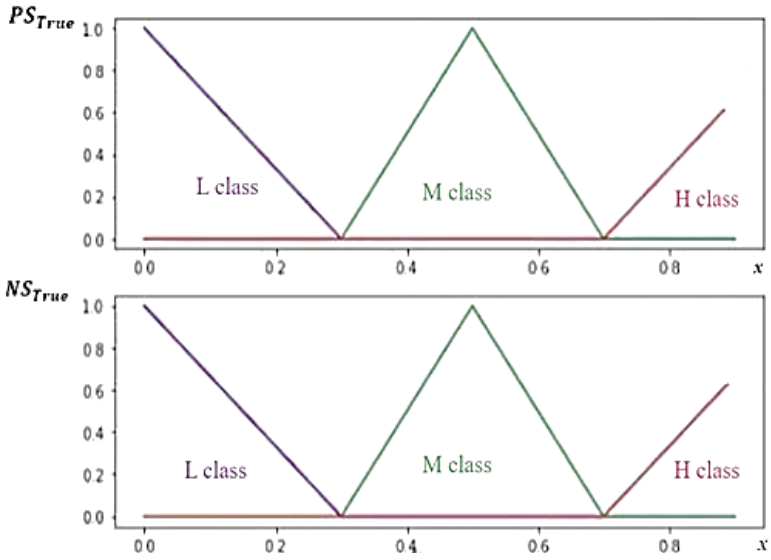


Figure 4. NL truth membership functions for the two inputs.

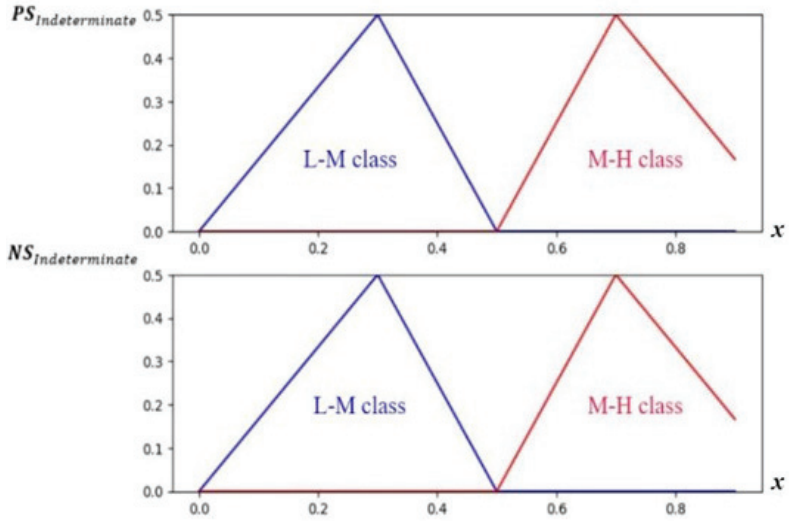


Figure 5. NL indeterminate membership functions for the two inputs.

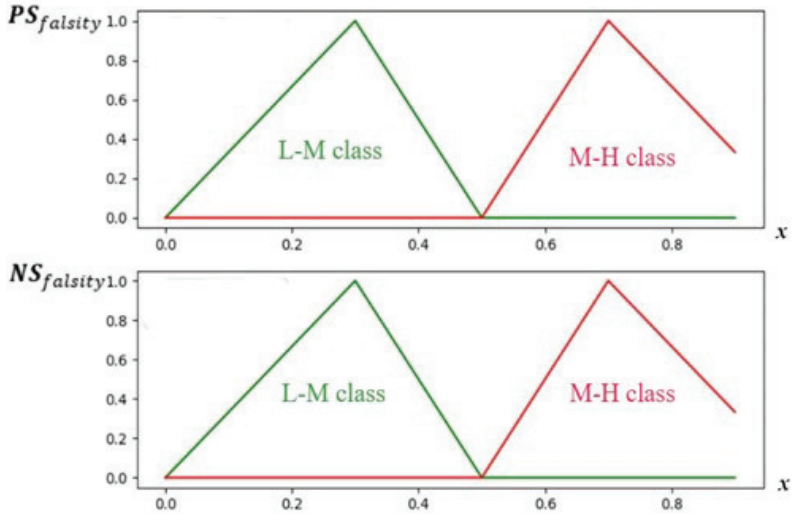


Figure 6. NL falsity membership functions for the two inputs.

(2) Rule Evaluation

In this phase, IF-THEN rules are produced by combining two inputs (i.e., PS and NS), each with levels. The rules are formulated for all truth, indeterminacy, and falsity inputs and outputs. The rules are inspired by the work carried out in [34]. The antecedent of the NL rules is represented by the inputs, each with three levels (i.e., L, M, H), and joined by the AND operator. Table 1 illustrates a sample of the NL rules.

Table 1. A sample of NL rules.

Rules	PS	NS	Output
R1	L	L	Neutral
R2	M	L	Positive
R3	L	M	Negative
R4	H	L	Positive
R5	H	H	Neutral
R6	M	H	Negative
R7	M	M	Neutral
R8	H	M	Positive
R9	L	H	Negative
R10	L-M	L-M	Negative–Neutral
R11	M-H	M-H	Positive–Neutral
R12	L-M	M-H	Negative–Neutral
R13	M-H	L-M	Positive–Neutral

(3) Deneutrosophication

In the final phase of our NL model, the neutrosophic output is transformed into a crisp output based on *COA* (center of the area) [33]. The output is calculated as follows:

$$COA = \frac{\sum z\mu_A(z)}{\sum \mu_A(z)} \tag{4}$$

where z represents the output variable, described in the neutrosophication step, while μ_A denotes the aggregated output of the rules. The neutrosophic result is obtained by combining the truth, indeterminacy, and falsity component values and is represented in the triplet format of (T, I, F). If the output falls within the overlapping zone of two membership functions, ambiguity is generated [12]. NL can handle ambiguity by defining a confidence value for the truth component according to Equation (5). If the truth value surpasses the confidence value (i.e., 0.5), the final output is the truth component value and the falsity and indeterminacy components are not significant; otherwise, the ambiguous results are generated [12,29]. Figure 7 illustrates how to determine the confidence value.

$$i|f = \begin{cases} \text{significant}, & t < 0.5 \\ \text{insignificant}, & t \geq 0.5 \end{cases} \tag{5}$$

where t, i, and f are the truth, indeterminacy, and falsity components, respectively. The final polarity is calculated from the neutrosophic output according to polarity classes in [34]. We have added a new polarity class called Indeterminate to the existing classes (e.g., Positive, Neutral, Negative) that is undecided, and we do not know whether it is negative, positive, or neutral according to Equation (2).

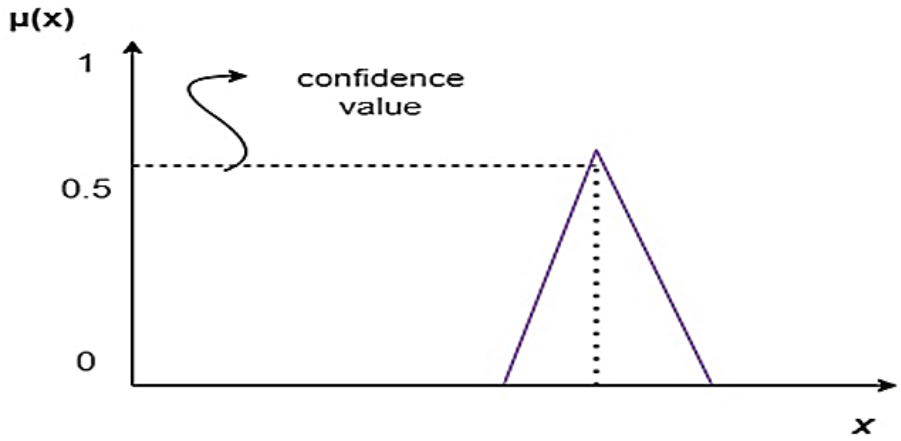


Figure 7. Determining the confidence value.

3.4. Prediction Model Using Long Short-Term Memory (LSTM)

This process is the last phase of our work. It is focused on predicting stock market movements by combining the sentiment polarity from the previous phase and historical stock price data obtained from the first phase to be fed into the prediction model. To accomplish this, we employ LSTM, a type of recurrent neural network that enables the preservation of input data information over extended periods [13]. The LSTM model comprises three layers, including an input layer, multiple LSTM layers, and finally, a single dense layer that consolidates the inputs received from the LSTM layer to generate the ultimate prediction value. Figure 8 depicts an LSTM architecture and its three corresponding layers.

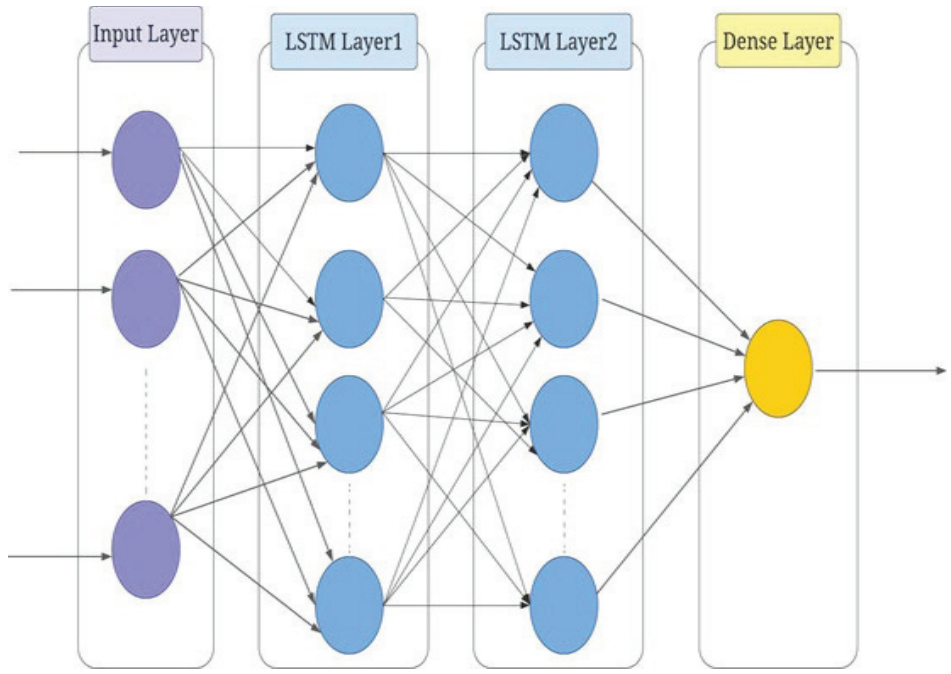


Figure 8. LSTM architecture.

The model has two hidden LSTM layers which improve the model's capacity to abstract information and depict more intricate patterns [13]. Each LSTM layer contains multiple memory cells. The memory cell contains three gates: forget, input, and output. These gates are constituted by the sigmoidal layer, which is crucial in regulating the transmission of information, both entering and exiting the memory cell [36]. The forget

gate is responsible for discerning which information should be removed from the cell state. This process is aided by the sigmoid activation function. The forget gate generates an output value within the 0 to 1 range for each element within the prior cell state. An output of 1 indicates the intention to preserve the associated information, whereas an output of 0 signifies the intent to discard that information [37]. This representation of the forget gate is denoted in Equation (6). The input gate oversees the incorporation of new information into the cell state. The sigmoid activation function governs which values will be updated. Meanwhile, the tanh activation function generates a set of new candidate values represented as a vector, which could be added to the state according to Equations (7)–(9). The output gate determines the resulting output for each cell. The output value is determined by the cell's current state in conjunction with the most recently added data according to Equations (10) and (11) [38].

$$f_t = \sigma\left(X_t \times U_f + H_{t-1} \times W_f\right) \tag{6}$$

$$i_t = \sigma\left(X_t \times U_i + H_{t-1} \times W_i\right) \tag{7}$$

$$\widetilde{C}_t = \tanh(X_t \times U_c + H_{t-1} \times W_c) \tag{8}$$

$$C_t = f_t \times c_{t-1} + i_t \times \widetilde{C}_t \tag{9}$$

$$o_t = \sigma\left(X_t \times U_o + H_{t-1} \times W_o\right) \tag{10}$$

$$H_t = o_t \times tanh(C_t) \tag{11}$$

where σ is the sigmoid function, X_t denotes the input at the current timestamp, U_f, U_i, U_c and U_o are the weights of the inputs, H_{t-1} represents the previous hidden state, and W_f, W_i, W_c and W_o are the weights of the hidden state. The terms f_t is the forget gate, i_t is the input gate, and o_t is the output gate. Moreover, $\widetilde{C}_t$ stands for the candidate cell state at the current timestamp, C_t represents the cell state at the current timestamp, and H_t denotes the current hidden state.

The input data to the model incorporate a set of parameters, including the adjusted closing price, opening price, highest price, lowest price, volume, and sentiment polarity, which were derived from the prior phase. We performed a binary classification to predict the stock market movement on day d based on the integration of the SA scores with the historical stock market data within a predetermined lag window of D days, encompassing the period $[d - D, d - 1]$. As an illustration, employing a lag window of D = 5 days implies the inclusion of data from the preceding five days [14,21]. The stock market movement is obtained by using the following formula:

$$m_d = \begin{cases} 0, & p_d^c < p_{d-1}^c \\ 1, & p_d^c \geq p_{d-1}^c \end{cases} \tag{12}$$

where m_d represents the movement on day d, p_d^c indicates the adjusted closing price on day d, and p_{d-1}^c indicates the adjusted closing price on the previous day. Additionally, 0 represents a downward movement and 1 represents an upward movement.

We evaluated the model's classification performance using accuracy, and Matthew's Correlation Coefficient (MCC) as evaluation metrics according to previous studies on stock prediction [14,21]. MCC is calculated as follows:

$$MCC = \frac{t_p t_n - f_p f_n}{\sqrt{\left(t_p + f_p\right)\left(t_p + f_n\right)\left(t_n + f_p\right)\left(t_n + f_n\right)}} \tag{13}$$

where t_p, t_n, f_p and f_n are true positive, true negative, false positive, and false negative, respectively.

To make a trading decision, we evaluated the model's financial performance using two metrics: cumulative return, which captures the total profit or loss over the test period according to Equation (14); and Sharpe ratio, which assesses the efficacy of an investment relative to its associated risk according to Equation (15) [21,38].

$$return_d = \sum_{i \in S} \frac{p_i^d - p_i^{d-1}}{p_i^{d-1}} (-1)^{Action_i^{d-1}} \tag{14}$$

where S is the set of all stocks considered in the analysis, p_i^d refers to the price of a specific stock i on particular day d, and $Action_i^{d-1}$ indicates the investment action taken for stock i on the previous day $(d-1)$ with value 0 or 1. Specifically, a value of 0 signifies a long position, indicating the investor purchased the stock on day $d-1$ with the anticipation of future price appreciation. Conversely, a value of 1 denotes short position action.

$$Sharpe\ Ratio_a = \frac{R_a - R_f}{\sigma_a} \tag{15}$$

where R_a represents the return, R_f indicates the risk-free rate, and σ_a represents the standard deviation of R_a.

4. Experimental Results

In this section, the experimental results and implementation of the proposed model are described. We have implemented our model in Python version 3.8.10. Due to the lack of an NL toolbox, scikit-fuzzy, a Python package for fuzzy logic, was opted for to implement NL by using three fuzzy inference systems to correspond to the truth, indeterminacy, and falsity components, as it was previously indicated in the previous studies that NL could be implemented using the fuzzy toolbox [12,29,32]. Furthermore, we selected the Keras Python library for implementing LSTM.

We utilized the StockNet dataset benchmark for tweets and historical stock market data to train and test our model [14,21]. Our dataset was split into training and testing sets in a ratio of 80:20. We shifted a window of 5 days for constructing input samples. Our LSTM model consists of one input layer, two LSTM layers, and a dense layer. We used the Sigmoid and Tanh as activation functions for the two LSTM layers. We trained the model for 10,000 epochs, using early stopping based on the MCC metric on the validation set to prevent overfitting [14]. We compared our work with the other models that use the same dataset to assess the effectiveness of our model [14,21].

4.1. Comparison of the Effect of Using Sentiment Analysis on the Prediction Model

The primary objective of the first set of experiments is to ascertain whether the sentiment of Twitter data has a significant effect on the performance of our prediction model, which was implemented using LSTM. The model was trained and evaluated on two datasets: one containing only historical stock market data; and the other containing a fusion of historical stock market data and sentiment scores obtained from Twitter related to users' feedback and opinions about specific stocks and companies, identified by their ticker symbols (e.g., $MSFT for Microsoft). SA was conducted using our NL model. As illustrated in Table 2, our model using sentiments based on NL achieves the highest performance, with an accuracy of 78.48% and an MCC score of 0.587. The results indicate that the model incorporating sentiment features outperforms the model without sentiment across all analyzed stocks, leading to a statistically significant improvement in accuracy. This confirms that the expression of positive or negative feedback by users about a given company on Twitter influences future changes in stock prices. Specifically, when the sentiment score is positive, we observe an increase in stock prices in the following days. This highlights a strong correlation between the public sentiment expressed on Twitter and the subsequent movement of stock prices, emphasizing that to obtain a highly accurate predictive model, a diverse range of data sources is required because using historical stock market data only

may have limitations in comprehensively encapsulating all pertinent variables influencing stock prices. In contrast, sentiment data can provide valuable insights into the market that are not readily discernible within conventional financial datasets.

Table 2. Comparison of the model's performance using sentiment based on NL, and without sentiment analysis.

Model	Accuracy	MCC
With sentiment using NL	78.48%	0.587
Without sentiment	73.28%	0.461

4.2. Comparison between Different Sentiment Analysis Techniques

The second set of experiments was conducted to assess the effectiveness of the machine learning sentiment classifier (i.e., NL model) in the SA process compared to the lexicon-based sentiment classifier used in [26] (i.e., TextBlob in our case) in our prediction model. We aggregated the result of SA with the historical stock market data and fed it into the LSTM model. The results in Table 3 show that SA using NL achieves the highest accuracy of around 78.48% and an MCC score of 0.587. Meanwhile, TextBlob achieves an accuracy, 75%, and an MCC score of 0.5119. The model prediction with NL-based SA yields a better result due to the ability of NL to classify the tweets and assess its ability to handle classification ambiguity because it assigns each instance true, indeterminate, and false values that accurately reflect the possible classifications of that instance. This leads to improved SA results, which, in turn, has a significant impact on the performance of our prediction model.

Table 3. Comparison of the model's performance using sentiments based on NL, and with sentiments using TextBlob.

Model	Accuracy	MCC
With sentiments using NL	78.48%	0.587
With sentiments using TextBlob	75%	0.5119

4.3. Comparison between Different Machine Learning Models

The third set of experiments was conducted to investigate the performance of the prediction model using LSTM compared to other prediction models using different machine learning techniques used in the previous studies such as naïve Bayes, neural networks, and Support Vector Machine [20,26]. The models were trained and evaluated using the same dataset containing a fusion of historical stock market data and sentiment scores. The results in Table 4 show that the LSTM model outperformed the other models in predicting the movement of the stock market. The LSTM model achieved the best result due to its proficiency in capturing long-term dependencies and its capacity to retain and leverage historical information for future predictions, a crucial advantage considering the influence of historical trends and social media sentiment on stock prices.

Table 4. Comparison of the model's performance using LSTM, neural networks, Support Vector Machine, and naïve Bayes.

Model	Accuracy	MCC
LSTM prediction model	78.48%	0.587
NN prediction model	72.4%	0.449
Support Vector Machine prediction model	62.06%	0.228
Naïve Bayes prediction model	58%	0.139

4.4. Comparison between Different Baseline Models

The fourth set of experiments aimed to ascertain the efficiency of the proposed model in comparison to the below baseline models employed in previous studies using the StockNet dataset.

- StockNet: a deep generative model employing historical data and twitter data to predict the stock market movement [14].
- Multipronged Attention Network for Stock Forecasting (MAN-SF): this model employs a joint deep learning architecture that integrates historical data, twitter data, and inter-stock correlations to predict the stock market movement [21].
- Adversarial Attentive LSTM: a deep learning model comprising four layers, incorporating Adversarial Training to emulate the stochastic nature of the stock price variable during the training process, thereby improving the accuracy of stock market predictions [39,40].

We opted for a black-box approach, utilizing the default configurations of baseline models. As shown in Table 5, our proposed model outperforms the others, exhibiting superior accuracy and MCC scores, while the Adversarial LSTM model achieved the worst performance in accuracy and MCC scores. These results demonstrate that our model is more effective because it factors in both the SA score and the historical stock market data, while the Adversarial LSTM model only uses historical stock market data. This highlights the ability of our model to yield the best result by using NL in SA to handle uncertainty and incomplete data in tweets, a facet not addressed by other models in the SA process. Additionally, our utilization of an LSTM model for stock market prediction, as previously noted, resulted in a notably high level of accuracy.

Table 5. Comparison of the performance between different baseline models.

Model	Accuracy	MCC
StockNet model	58.2%	0.081
Adversarial Attentive LSTM model	57.2%	0.148
MAN-SF model.	60.8%	0.195
Our LSTM model with sentiments using NL.	78.48%	0.587

4.5. Comparison of the Financial Performance

To make a trading decision and assess the financial performance, we conducted a set of experiments to compare our model with other baselines and a benchmark strategy of buy and hold. As shown in Table 6, we compared our model with other models employed in previous studies using the same dataset [14,21]. Our model demonstrated superior performance compared to various established models, exhibiting superior return and Sharpe ratio scores (similar Sharpe ratio score to MAN-SF). These results indicate strong potential for generating excess return and managing risk. Furthermore, the findings underscore the profitability of our model when contrasted with these models that do not incorporate NL in SA process. Furthermore, a comparison was made between our proposed model and the buy-and-hold strategy, as detailed in Table 7. Our proposed model demonstrated a superior performance, surpassing the buy-and-hold strategy in terms of return and Sharpe ratio scores. This suggests the potential of our model to enhance profitability in financial forecasting.

Table 6. Comparison of the financial performance between different models.

Model	Return	Sharpe Ratio
StockNet	0.9%	0.83
MAN-SF	1.66%	1
Our proposed model	4.1%	1

Table 7. Comparison of the financial performance between our proposed model and buy-and-hold strategy.

Metrics	Buy and Hold	Our Proposed Model
Return	0.6%	4.1%
Sharpe Ratio	0.3	1

5. Conclusions and Future Work

In this paper, we introduced a stock market movement prediction model that fuses social media data with historical stock price data. Due to the importance of stock market prediction in recent years and the significant influence of social media on it, the problem of data uncertainty and ambiguity in social media decreases the accuracy of SA results, thereby reducing the accuracy of the stock market forecasting model that utilizes SA. Previous studies did not address this problem. Therefore, the main purpose of our work is to enhance the performance of stock movement prediction by improving the SA results of the tweets through the utilization of NL integrated with a lexicon-based approach capable of handling ambiguous, incomplete, and uncertain data collected from Twitter, particularly in relation to individuals' perspectives on corporations and stocks and companies. Our proposed model demonstrated its advantage by utilizing the StockNet dataset benchmark and comparing it to models that use this dataset. The proposed model feeds the integrated SA scores with historical stock market data into an LSTM model to foresee the stock movement. Notably, our model distinguishes itself as the first to employ NL in the SA process to predict stock market movement. The findings highlight the importance of incorporating social media data into stock market prediction models because it showed superiority in accuracy and MCC score in comparison to the model without the integration of SA. The proposed model outperformed other models that utilized the same dataset by utilizing NL in the SA process to make the results more compatible with human sentiment and using the integration of historical stock market data with SA results as input factors to our prediction model using LSTM, which resulted in a relatively high accuracy, of around 78.48%, and an MCC score of 0.587. Our investigation into the impact of the NL model in SA on prediction performance revealed that it outperformed the model that utilized TextBlob in the SA. Furthermore, we conducted an examination of the efficiency of employing LSTM in our prediction model, finding that it outperforms models using naïve Bayes, neural networks, and Support Vector Machine. In our model, we also measured the performance of our model based on financial metrics by calculating the return and Sharpe ratio to help investors in trading decisions. We compared our model with baselines and the buy-and-hold strategy. The results showed the superiority of our model in return and Sharpe ratio scores, indicating strong potential for generating excess return and managing risk. In the future, we aim to consider the effect of the variety of user profiles, in terms of whether they are experts, investors, qualified, influencers, or students, on the SA results and seek to use other social media platforms' data, such as Facebook, LinkedIn, and StockTwits. Additionally, we intend to detect and filter out spam tweets and posts to enhance the sentiment classification performance. Moreover, we aim to utilize a larger benchmark dataset in comparison to our dataset.

Author Contributions: Conceptualization, S.M.D. and B.A.A.; methodology, S.M.D.; software, B.A.A.; validation, S.M.D., S.M.E. and B.A.A.; formal analysis, S.M.D.; investigation, S.M.E.; resources, B.A.A.; data curation, B.A.A.; writing—original draft preparation, B.A.A.; writing—review and editing, S.M.D.; visualization, B.A.A.; supervision, S.M.D. and S.M.E.; project administration, S.M.D. and S.M.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xiao, Q.; Ihnaini, B. Stock trend prediction using sentiment analysis. *PeerJ Comput. Sci.* **2023**, *9*, e1293. [CrossRef] [PubMed]
2. Khan, W.; Ghazanfar, M.A.; Azam, M.A.; Karami, A.; Alyoubi, K.H.; Alfakeeh, A.S. Stock market prediction using machine learning classifiers and social media, news. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *13*, 3433–3456. [CrossRef]
3. Ruan, Y.; Durresi, A.; Alfantoukh, L. Using Twitter trust network for stock market analysis. *Knowl.-Based Syst.* **2018**, *145*, 207–218. [CrossRef]
4. Gumus, A.; Sakar, C.O. Stock Market Prediction by Combining Stock Price Information and Sentiment Analysis. *Int. J. Adv. Eng. Pure Sci.* **2021**, *33*, 18–27. [CrossRef]
5. Beg, M.O.; Awan, M.N.; Ali, S.S. Algorithmic Machine Learning for Prediction of Stock Prices. In *FinTech as a Disruptive Technology for Financial Institutions*; IGI Global: Hershey, PA, USA, 2019; pp. 142–169.
6. Chandola, D.; Mehta, A.; Singh, S.; Tikkiwal, V.A.; Agrawal, H. Forecasting Directional Movement of Stock Prices using Deep Learning. *Ann. Data Sci.* **2022**, *10*, 1361–1378. [CrossRef]
7. Mankar, T.; Hotchandani, T.; Madhwani, M.; Chidrawar, A.; Lifna, C.S. Stock market prediction based on social sentiments using machine learning. In Proceedings of the International Conference on Smart City and Emerging Technology, Mumbai, India, 5 January 2018; pp. 1–3.
8. Rajendiran, P.; Priyadarsini, P. Survival study on stock market prediction techniques using sentimental analysis. *Mater. Today Proc.* **2023**, *80*, 3229–3234. [CrossRef]
9. Colhon, M.; Vlăduţescu, Ş.; Negrea, X. How Objective a Neutral Word Is? A Neutrosophic Approach for the Objectivity Degrees of Neutral Words. *Symmetry* **2017**, *9*, 280. [CrossRef]
10. Kandasamya, I.; Vasanthaa, W.; Obbinenib, J.; Smarandache, F. Sentiment analysis of tweets using refined neutrosophic sets. *Comput. Ind.* **2020**, *115*, 103180–103190. [CrossRef]
11. AboElHamd, E.; Shamma, H.M.; Saleh, M.; El-Khodary, I. Neutrosophic logic theory and applications. *Neutrosophic Sets Syst.* **2021**, *41*, 4.
12. Essameldin, R.; Ismail, A.A.; Darwish, S.M. An Opinion Mining Approach to Handle Perspectivism and Ambiguity: Moving Toward Neutrosophic Logic. *IEEE Access* **2022**, *10*, 63314–63328. [CrossRef]
13. Heiden, A.; Parpinelli, R.S. Applying LSTM for stock price prediction with sentiment analysis. In Proceedings of the Fifteenth Brazilian Congress of Computational Intelligence, Online, 26–29 October 2021; pp. 1–8.
14. Xu, Y.; Cohen, S.B. Stock movement prediction from tweets and historical prices. In Proceedings of the Fifty-Sixth Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 1970–1979.
15. Asghar, M.Z.; Rahman, F.; Kundi, F.M.; Ahmad, S. Development of stock market trend prediction system using multiple regression. *Comput. Math. Organ. Theory* **2019**, *25*, 271–301. [CrossRef]
16. Kalyani, J.; Bharathi, P.; Jyothi, P. Stock trend prediction using news sentiment analysis. *Int. J. Comput. Sci. Inf. Technol.* **2016**, *8*, 67–76.
17. Pagolu, V.S.; Reddy, K.N.; Panda, G.; Majhi, B. Sentiment analysis of twitter data for predicting stock market movements. In Proceedings of the International Conference on Signal Processing, Communication, Power and Embedded System, Paralakhemundi, Odisha, India, 3–5 October 2016; pp. 1345–1350.
18. Xu, Y.; Keselj, V. Stock prediction using deep learning and sentiment analysis. In Proceedings of the IEEE International Conference on Big Data, Los Angeles, CA, USA, 9–12 December 2019; pp. 5573–5580.
19. Maqsood, H.; Mehmood, I.; Maqsood, M.; Yasir, M.; Afzal, S.; Aadil, F.; Selim, M.M.; Muhammad, K. A local and global event sentiment based efficient stock exchange forecasting using deep learning. *Int. J. Knowl. Manag.* **2020**, *50*, 432–451. [CrossRef]
20. Gupta, R.; Chen, M. Sentiment analysis for stock price prediction. In Proceedings of the 2020 IEEE Conference on Multimedia Information Processing and Retrieval, Shenzhen, China, 6–8 August 2020; pp. 213–218.
21. Sawhney, R.; Agarwal, S.; Wadhwa, A.; Shah, R. Deep attentive learning for stock movement prediction from social media text and company correlations. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online, 16–20 November 2020; pp. 8415–8426.
22. Ho, T.-T.; Huang, Y. Stock Price Movement Prediction Using Sentiment Analysis and CandleStick Chart Representation. *Sensors* **2021**, *21*, 7957. [CrossRef] [PubMed]
23. Fazlija, B.; Harder, P. Using Financial News Sentiment for Stock Price Direction Prediction. *Mathematics* **2022**, *10*, 2156. [CrossRef]
24. Cristescu, M.P.; Nerisanu, R.A.; Mara, D.A.; Oprea, S.-V. Using Market News Sentiment Analysis for Stock Market Prediction. *Mathematics* **2022**, *10*, 4255. [CrossRef]
25. Srijiranon, K.; Lertratanakham, Y.; Tanantong, T. A Hybrid Framework Using PCA, EMD and LSTM Methods for Stock Market Price Prediction with Sentiment Analysis. *Appl. Sci.* **2022**, *12*, 10823. [CrossRef]
26. Koukaras, P.; Nousi, C.; Tjortjis, C. Stock Market Prediction Using Microblogging Sentiment Analysis and Machine Learning. *Telecom* **2022**, *3*, 358–378. [CrossRef]
27. Costola, M.; Hinz, O.; Nofer, M.; Pelizzon, L. Machine learning sentiment analysis, COVID-19 news and stock market reactions. *Res. Int. Bus. Financ.* **2023**, *64*, 101881. [CrossRef]

28. Awajan, I.; Mohamad, M.; Al-Quran, A. Sentiment analysis technique and neutrosophic set theory for mining and ranking big data from online reviews. *IEEE Access* **2021**, *9*, 47338–47353. [CrossRef]
29. Ansaria, A.; Biswasb, R.; Aggarwal, S. Neutrosophic classifier: An extension of fuzzy classifier. *Appl. Soft Comput.* **2013**, *13*, 563–573. [CrossRef]
30. Kandasamy, I.; Vasantha, W.B.; Mathur, N.; Bisht, M.; Smarandache, F. Sentiment analysis of the #metoo movement using neutrosophy: Application of single-valued neutrosophic sets. In *Optimization Theory Based on Neutrosophic and Plithogenic Sets*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 117–135. [CrossRef]
31. Madbouly, M.M.; Darwish, S.M.; Essameldin, R. Modified fuzzy sentiment analysis approach based on user ranking suitable for online social networks. *IET Softw.* **2020**, *14*, 300–307. [CrossRef]
32. Essameldin, R.; Ismail, A.A.; Darwish, S.M. Quantifying Opinion Strength: A Neutrosophic Inference System for Smart Sentiment Analysis of Social Media Network. *Appl. Sci.* **2022**, *12*, 7697. [CrossRef]
33. Hassan, M.H.; Darwish, S.M.; Elkaffas, S.M. An Efficient Deadlock Handling Model Based on Neutrosophic Logic: Case Study on Real Time Healthcare Database Systems. *IEEE Access* **2022**, *10*, 76607–76621. [CrossRef]
34. Vashishtha, S.; Susan, S. Fuzzy rule based unsupervised sentiment analysis from social media posts. *Expert Syst. Appl.* **2019**, *138*, 112834. [CrossRef]
35. Abdel-Basset, M.; Gunasekaran, M.; Mohamed, M.; Smarandache, F. A novel method for solving the fully neutrosophic linear programming problems. *Neural Comput. Appl.* **2019**, *31*, 1595–1605. [CrossRef]
36. Ko, C.-R.; Chang, H.-T. LSTM-based sentiment analysis for stock price forecast. *PeerJ Comput. Sci.* **2021**, *7*, e408. [CrossRef]
37. John, A.; Latha, T. Stock market prediction based on deep hybrid RNN model and sentiment analysis. *Automatika* **2023**, *64*, 981–995. [CrossRef]
38. Moghar, A.; Hamiche, M. Stock Market Prediction Using LSTM Recurrent Neural Network. *Procedia Comput. Sci.* **2020**, *170*, 1168–1173. [CrossRef]
39. Kim, R.; So, C.H.; Jeong, M.; Lee, S.; Kim, J.; Kang, J. Hats: A hierarchical graph attention network for stock movement prediction. *arXiv* **2019**, arXiv:1908.07999.
40. Feng, F.; Chen, H.; He, X.; Ding, J.; Sun, M.; Chua, T.S. Enhancing stock movement prediction with adversarial training. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; pp. 5843–5849.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Journal of
Theoretical and Applied
Electronic Commerce Research

MDPI

Article

Financial Anti-Fraud Based on Dual-Channel Graph Attention Network

Sizheng Wei [1,2] and Suan Lee [2,*]

[1] School of Finance, Xuzhou University of Technology, Xuzhou 221018, China; wei4zheng@xzit.edu.cn
[2] School of Computer Science, Semyung University, Jecheon-si 27136, Republic of Korea
* Correspondence: suanlee@semyung.ac.kr; Tel.: +82-43-649-1273

Abstract: This article addresses the pervasive issue of fraud in financial transactions by introducing the Graph Attention Network (GAN) into graph neural networks. The article integrates Node Attention Networks and Semantic Attention Networks to construct a Dual-Head Attention Network module, enabling a comprehensive analysis of complex relationships in user transaction data. This approach adeptly handles non-linear features and intricate data interaction relationships. The article incorporates a Gradient-Boosting Decision Tree (GBDT) to enhance fraud identification to create the GBDT–Dual-channel Graph Attention Network (GBDT-DGAN). In a bid to ensure user privacy, this article introduces blockchain technology, culminating in the development of a financial anti-fraud model that fuses blockchain with the GBDT-DGAN algorithm. Experimental verification demonstrates the model's accuracy, reaching 93.82%, a notable improvement of at least 5.76% compared to baseline algorithms such as Convolutional Neural Networks. The recall and F1 values stand at 89.5% and 81.66%, respectively. Additionally, the model exhibits superior network data transmission security, maintaining a packet loss rate below 7%. Consequently, the proposed model significantly outperforms traditional approaches in financial fraud detection accuracy and ensures excellent network data transmission security, offering an efficient and secure solution for fraud detection in the financial domain.

Keywords: financial anti-fraud; graph neural networks; graph attention network; deep learning; blockchain

1. Introduction

In the rapidly evolving landscape of finance, thwarting fraudulent activities has emerged as a paramount concern for both financial institutions and regulatory bodies. Traditional fraud detection methods, relying heavily on personal information and historical records, face significant limitations in the face of technological advancements and the escalating complexity of financial transactions. These limitations are particularly pronounced when dealing with collaborative fraudulent activities involving multiple decisions and dynamically changing, high-dimensional data [1,2]. In today's digital age, the rapid growth of financial transactions and the continuous evolution of financial fraud have led to an urgent need for more innovative and efficient anti-fraud solutions. With the development of financial technology, malicious actors have become increasingly cunning, using more complex means to engage in fraudulent activities, posing a serious threat to the stability of the financial system and the security of users' assets.

So, in the face of increasingly complex financial transactions and technological advancements, how do we improve the generalization ability of financial anti-fraud models, ensure the effectiveness of the models, and protect customer privacy? The financial industry is thus confronted with an urgent need for more robust and intelligent fraud detection tools. In response to this challenge, deep learning, an artificial intelligence algorithm, has emerged as a promising solution. Through unsupervised learning on data information, deep learning can adeptly navigate complex financial network structures, comprehensively

discerning data patterns and identifying potential abnormal behaviors or fraud patterns. This capability opens up new avenues for effectively addressing the intricate issue of financial fraud [3].

The prevalence of fraudulent activities in financial transactions has escalated into a pervasive global challenge. Graph Neural Networks (GNNs), serving as deep learning models designed for graph-structured data, demonstrate proficiency in modeling intricate financial networks encompassing customer relationships, transaction processes, and fund flows. By leveraging their capacity to holistically capture patterns and regularities within data, GNNs excel in identifying potential abnormal behaviors and fraud patterns embedded in financial datasets. Their adeptness in handling unstructured data and discerning complex relationships between nodes provides them with a distinct advantage in uncovering potential risks and abnormal transactions [4].

Traditional anti-fraud methods have difficulty in effectively capturing these new types of fraudulent behaviors, so researchers are constantly striving to seek innovative solutions based on advanced technologies. A comprehensive survey by Ma et al. (2021) delved into deep learning methods in graph anomaly detection, underscoring the application advantages of GNN models in detecting anomalies within graph data [5]. Ashtiani and Raahemi (2021) compiled a comprehensive summary of applications and methodologies related to deep learning algorithms, including GNN-based works, in fraud detection. Their work offers an in-depth exploration and outlines research directions within the realm of financial fraud detection [6]. Additionally, Li et al. (2021) proposed a medical image fusion method based on deep learning, presenting more accurate and comprehensive image data prediction results for medical diagnosis and treatment [7].

Despite the considerable promise offered by GNNs in the realm of financial anti-fraud, practical implementations face several formidable challenges. Issues such as data quality and scale, model interpretability, and privacy protection emerge as significant hurdles. To address these challenges, blockchain technology, renowned for its decentralized, tamper-resistant, and encryption security features, emerges as a powerful solution, particularly in enhancing the safeguarding of financial data privacy. By adopting a decentralized and distributed approach to storing transaction data, blockchain serves as a robust deterrent against data tampering and unauthorized access. This innovative application of technology empowers financial institutions to securely share and verify data, all while ensuring comprehensive protection of customer privacy [8,9].

In conclusion, the goal of this article is to improve the generalization ability of financial anti-fraud models while ensuring data privacy, in response to the challenges posed by increasingly complex fraudulent activities in the financial sector. This article first collects financial transaction data and performs preprocessing and feature engineering to provide input for the GNN model. Second, this article designs and trains a GNN model to identify patterns and anomalies in the data through unsupervised learning, thus identifying potential fraudulent patterns and abnormal behaviors. Next, this article integrates blockchain technology into the data storage and sharing process to ensure the privacy and security of transaction data. Finally, rigorous model performance evaluation and comparative experiments are conducted to propose improvements and solutions to real-world challenges. The contribution lies in providing financial institutions with more powerful fraud detection tools, promoting the development of financial technology, and promoting sustainable growth and secure operation of the financial industry. The synergistic application of GNN and blockchain heralds a new era in financial security, promising heightened efficiency and resilience against evolving threats in the dynamic landscape of financial transactions.

2. Literature Review

2.1. Review of Intelligent Recognition Research in Financial Anti-Fraud

In the dynamic landscape of financial fraud, traditional rules and statistical models encounter challenges in adapting to evolving methodologies, necessitating the exploration of more flexible and adaptive approaches. Many scholarly investigations have contributed

significantly to the field of financial anti-fraud. Liu and Jiang (2023) [10] introduced an advanced persistent threat prediction analysis method, employing causal graph modeling to forecast advanced persistent threats—a pivotal stride in predictive analysis. Fang et al. (2021) [11] innovatively addressed internet loan fraud by integrating deep learning techniques, enhancing accuracy and efficiency in detecting fraudulent activities on online lending platforms. Alarfaj et al. (2022) [12] dedicated their efforts to employing advanced machine learning and deep learning algorithms in credit card fraud detection. Their model harnessed cutting-edge technologies, heightening the accuracy of credit card transaction fraud detection systems and fortifying the security of financial transactions. Cheng et al. (2020) [13] proposed a credit card fraud detection neural network integrating a spatiotemporal attention mechanism. This attention mechanism effectively captured spatiotemporal patterns in credit card transactions, augmenting the precision and reliability of fraud detection in financial transactions. Lei et al. (2022) [14] ventured into the development of an intelligent information system for financial analysis, leveraging supervised machine learning algorithms. This system facilitated comprehensive financial analysis, aiding decision-making by integrating supervised machine learning technology. In a consumer-centric approach, Xu et al. (2022) [15] constructed a consumer fraud victim model by incorporating individual-specific factors. This enhancement significantly improved the accuracy of the fraud victim prediction model, contributing to the construction of a more refined consumer fraud detection system. In pursuing enhanced privacy protection, Verykios et al. (2022) [16] delved into techniques for concealing sensitive data in the financial anti-fraud process. Their focus on safeguarding sensitive data during anti-fraud procedures maintained privacy while preserving an effective fraud detection mechanism. Exploring graph-based machine learning, Usman et al. (2023) [17] proposed an intelligent anti-money laundering fraud control model. Leveraging graph-based machine learning technology, their model elevated anti-money laundering measures, consequently enhancing fraud control and compliance in financial transactions. Collectively, these studies underscore the multifaceted and evolving nature of intelligent recognition in the realm of financial anti-fraud research.

2.2. Comprehensive Review of GNN Applications in the Financial Landscape

GNNs have emerged as powerful tools for handling unstructured financial data, offering a holistic approach to capturing intricate financial network structures. The versatility of GNNs in financial research is evident through various studies conducted by scholars in the field. Cheng et al. (2022) [18] proposed a multimodal Graph Attention Network (GAN) for financial time series prediction. Integrating diverse data modalities into a graph structure, the model significantly enhances the accuracy of financial time series analysis. By harnessing the power of GNNs, the model adeptly captures intricate relationships within financial data, thereby improving predictive performance. In a distinct domain, Zhang et al. (2023) [19] introduced a dynamic attribute-driven GAN for stock prediction in behavioral finance. By leveraging dynamic attributes to construct a GAN, this model effectively captures the evolving relationships among stock behaviors, considering dynamic features within the financial market to enhance the accuracy of stock prediction. Xu et al. (2022) [20] proposed a Hierarchical GAN tailored for classifying stocks with price limit changes. Embracing a hierarchical structure, the model adeptly captures multi-level relationships within financial data, thereby refining the accuracy of classifying stocks subject to price limit changes. Lazcano et al. (2023) [21] presented a combined model integrating recurrent neural networks and graph convolutional networks for financial time series prediction. This amalgamation effectively captures both time dependencies and graph-based features in financial data, resulting in improved accuracy for financial time series prediction. In conclusion, these studies underscore the versatility and efficacy of GNNs in various financial domains, showcasing their potential to enhance accuracy, capture intricate relationships, and refine predictive capabilities. Sun et al. (2024) [22] utilized GraphSAGE and deep reinforcement learning for optimizing financial investment portfolios. The results

showed that this method performed well in handling complex financial relationships, but more empirical research was needed to verify its effectiveness in actual markets.

In summary, a meticulous examination of the reviewed literature underscores the substantial potential of emerging technologies, particularly machine learning and deep learning, in the realm of fraud detection. Pioneering studies by Fang et al. (2021), Xu et al. (2022), Verykios et al. (2022), and Usman et al. (2023) illustrate methodologies adept at discerning intricate patterns within vast, high-dimensional datasets. These innovative approaches identify anomalous behavior effectively and exhibit adaptability to evolving forms of fraud through continuous optimization. While the exploration of graph neural network applications in finance highlights their remarkable efficacy in fraud recognition and stock prediction, the existing body of research has overlooked concerns related to the leakage of user data information within the financial domain.

Although the aforementioned research extensively explores the applications of machine learning and deep learning technologies in the financial field, there is little mention of the potential of GNNs in financial anti-fraud. And the application of graph neural networks in the field of financial fraud has not been fully explored. This article aims to use GNNs to solve the problem of anti-fraud in the financial field and improve the accuracy and reliability of fraud detection. In particular, this article focuses on the complex correlations and unstructured data of financial transactions and explores these characteristics through GNN technology to fill the gap in existing research in this field and provide new research perspectives and solutions for the field of financial anti-fraud.

3. GNN Applications in Financial Anti-Fraud Research

3.1. Optimization Analysis of GNN Models Applied in Financial Anti-Fraud

In the realm of finance, intricate relationships among nodes in financial network data necessitate a nuanced approach for effective fraud detection. Recognizing the multifaceted nature of financial fraud requires a thorough understanding and capture of node information within financial network data. GNNs emerge as a powerful tool for this purpose, excelling in the capacity to capture associations between nodes and model complex relationships inherent in financial transactions [23,24]. It is noteworthy that GNN models encompass Graph Convolutional Neural Networks (GCNs) and GANs.

GCNs, rooted in convolutional operations, are tailored to handle graph data, specifically representing connections between nodes [25]. The fundamental concept of GCNs involves information propagation based on the underlying graph structure, facilitating each node in acquiring and integrating information from its neighboring nodes, as illustrated in Figure 1.

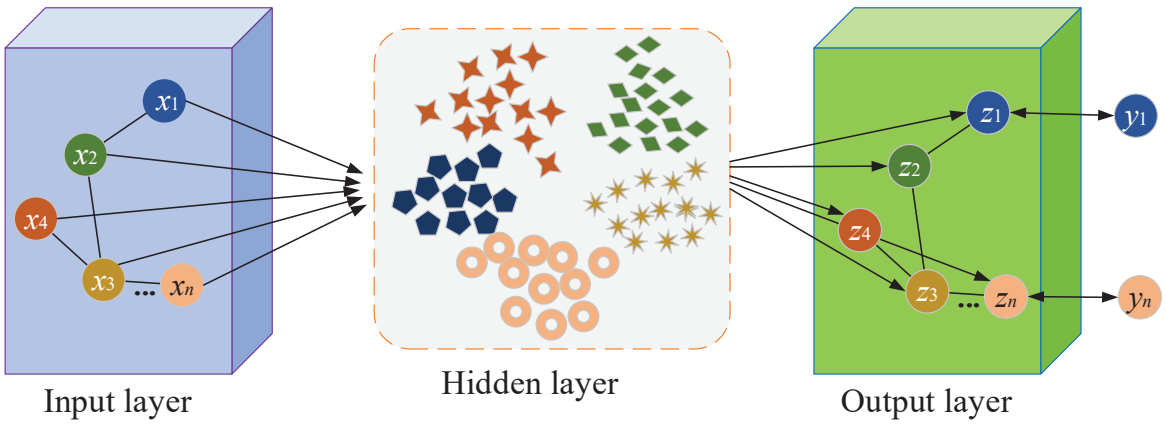


Figure 1. GCN framework.

In GCNs, each node aggregates information about its neighboring nodes and adjusts weights to retain important information about node relationships. After aggregating

neighbor information, nodes will propagate their feature information to neighboring nodes, forming a layer-by-layer information transmission process. At each layer, through graph convolution operations, nodes update their feature representations while considering the contributions of neighboring nodes. This graph-based information propagation enables GCNs to capture nonlinear relationships and interactions between nodes in complex networks, providing powerful modeling capabilities for the financial anti-fraud model.

Let the graph representing image information in the financial domain be denoted as $G = (V, E)$, where the graph comprises N vertices $v_i \in V$, with edge sets $e_{ij} \in (v_i, v_j) \in E$ and edge weight w_{ij}. First, vertex features $X = \{x_i\}_{i=1}^{N}$ v are extracted, where x_i denotes the feature vector of the i-th vertex v_i in the image. These features are then utilized as input for the GCN. In the context of the anti-fraud graph in the financial domain, x_i can be defined as in Equation (1) below:

$$x_i = concat\big(M_{AB}(v_i), M'_{AB}(v_i)\big) \tag{1}$$

Equation (2) describes the GCN.

$$H^{(l+1)} = \sigma\left(\overline{D}^{-\frac{1}{2}}\overline{AD}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{2}$$

In Equation (2), $H^{(l)}$ denotes the l-th hidden layer; $\overline{A} = I_N + A$ refers to the adjacency matrix of the subject interaction graph; I_N signifies the identity matrix; $\overline{D}_{ii} = \sum_j \overline{A}_{ij}$ refers to the degree matrix containing the entry and exit coefficients of each vertex of the interaction graph; $W^{(l)}$ represents the parameter matrix of network learning; and σ indicates the activation function, utilizing the Rectified Linear Unit (ReLU) activation.

The GAN accomplishes feature propagation and aggregation by adaptively learning attention weights between each node and its neighboring nodes [26,27].

In a standard GAN, the input and output of the graph attention layer are defined as in Equations (3) and (4), respectively:

$$h_{In} = \left\{\vec{h}_1, \vec{h}_2, \cdots, \vec{h}_n\right\}, h \in R^{n \times InD} \tag{3}$$

$$h_{Out} = \left\{\vec{h}_1, \vec{h}_2, \cdots, \vec{h}_n\right\}, h \in R^{n \times OutD} \tag{4}$$

Here, n stands for the number of vertices contained in the graph data, InD refers to the feature dimension of the input vertex, and $OutD$ represents the feature dimension of the output vertex.

For the target node q, a forward propagation neural network is utilized to calculate the weight e_{qu} of its neighboring nodes $N(q)$, as defined in Equation (5):

$$e_{qu} = a\big(\big[Wh_q \big|\big| Wh_u\big]\big),\ u \in N(q) \tag{5}$$

In Equation (5), $||$ refers to the vector-splicing operation, W denotes the learnable parameter matrix, and $a(\cdot)$ is the mapping function. h_u and h_q refer to the feature vectors. After obtaining the weights of neighboring nodes using GAN, the neighborhood weight α_{qu} is normalized using the softmax activation function, as defined in Equation (6):

$$\alpha_{qu} = softmax\big(e_{qu}\big) = \frac{exp\big(LeakReLU\big(e_{qu}\big)\big)}{\sum_{k \in N(q)} exp\big(LeakReLU\big(e_{qu}\big)\big)} \tag{6}$$

Once GAN computes the weight distribution α_{qu} for all neighboring nodes of the node q, it performs a weighted sum of the features of each neighboring node along with its weight. The feature vector h'_q of node q is updated, as shown in Equation (7):

$$h'_q = \sigma\left(\sum_{u \in N(q)} \alpha_{qu} W h_u\right) \tag{7}$$

Furthermore, GAN employs K independent attention heads to model relationships between nodes in different subspaces. The updated feature vector h'_q of node q is represented as Equation (8):

$$h'_q = \Big\|_{K}^{k=1} \sigma\left(\sum_{u \in N(q)} \alpha_{qu}^{k} W^{k} h_u\right) \tag{8}$$

The output of multi-head attention is formed by concatenating the features from all heads, providing a more comprehensive representation of node features in the graph. Additionally, the output of the last layer (prediction layer) in GAN is the average result across multiple attention heads, as denoted in Equation (9):

$$z_q = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{u \in N(q)} \alpha_{qu}^{k} W^{k} h_u\right) \tag{9}$$

Based on the aforementioned computation process of GAN, the node feature-updating process is illustrated in Figure 2.

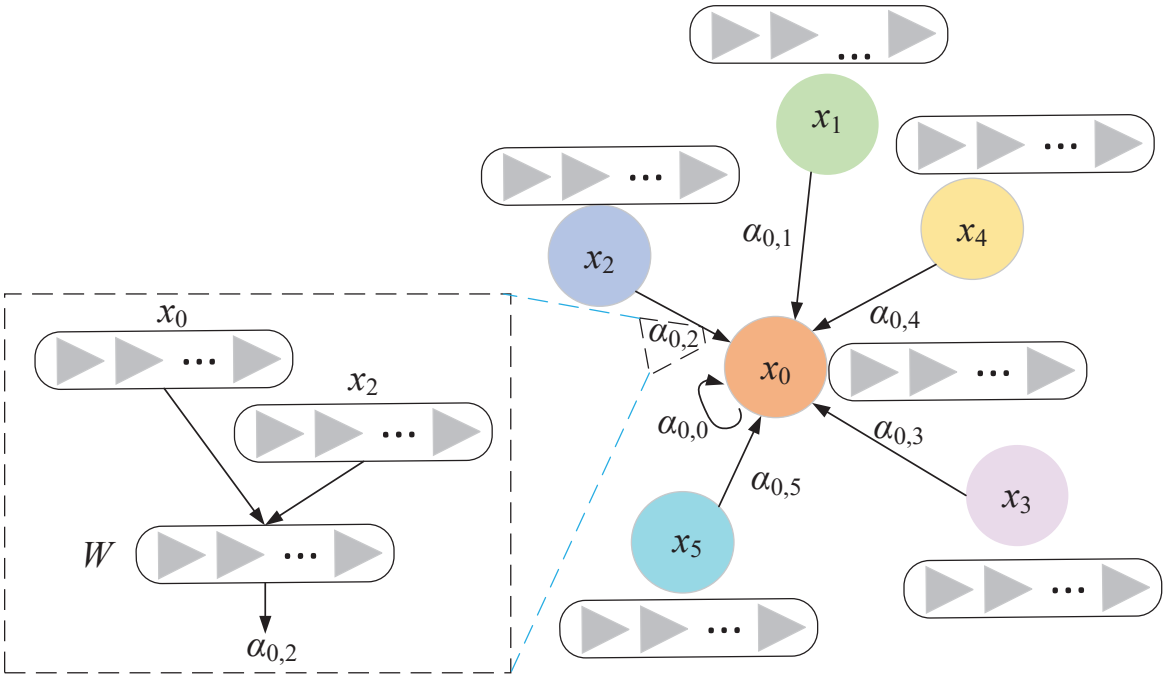


Figure 2. Node-updating process of GAN.

In Figure 2, within the GAN, the node feature updating involves the following steps. Step 1: Employ a learnable function to compute the "similarity" weights between the features of neighboring nodes and the target node's feature, followed by normalization using an activation function to form the weight distribution of neighboring nodes. Step 2: Weigh and sum the feature vectors of neighboring nodes to obtain a comprehensive feature vector of neighboring nodes. Step 3: Merge this integrated feature vector with the

feature vector of the target node itself. Step 4: Update the target node's feature through a non-linear transformation.

In this article, to mitigate redundancy and noise in fraud data within the financial domain network, a graph attention mechanism filters out noise and more accurately captures the semantic relationships between features and nodes. Simultaneously, the node channel attention and semantic channel attention in the GAN are combined to construct a Dual-channel Graph Attention Network (DGAN) module. This algorithm can filter out noise information, making the model more focused on extracting key features, thereby improving the sensitivity and accuracy of the model to data features. The design of the Dual-channel Attention Network module is illustrated in Figure 3.

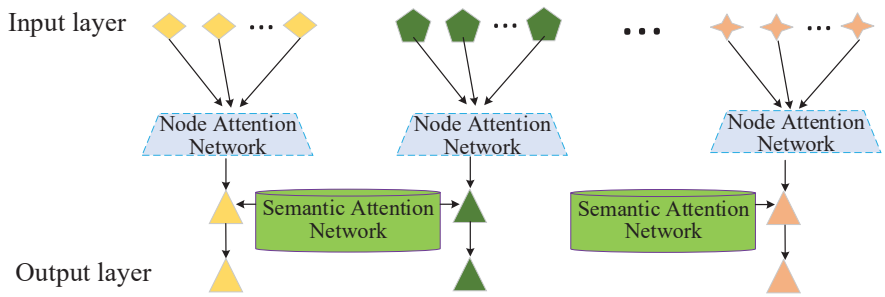


Figure 3. DGAN module.

Moreover, to adeptly manage non-linear features and intricate interactive relationships within the data, Gradient-Boosting Decision Tree (GBDT) [28] is incorporated. GBDT, as a powerful nonlinear model, has significant advantages in handling nonlinear features and complex data interactions. The DGAT network, which combines the graph attention mechanism, captures the relationship features between nodes more finely. GBDT is seamlessly integrated with the DGAN network, giving rise to the GBDT-DGAN network. The advantage of this combination lies in the fact that the GBDT-DGAT network can fully utilize the powerful nonlinear feature extraction ability of GBDT, while improving sensitivity to node relationships through the graph attention mechanism, thus considering the importance of features and relationships more comprehensively in the recognition of data information in the financial field. Subsequently, this amalgamated network is applied to identify data information within the financial domain.

3.2. Analysis of Privacy Protection with Blockchain

In the realm of financial transactions, ensuring data privacy holds paramount significance. Blockchain technology, distinguished by its decentralized, highly secure, and tamper-resistant distributed ledger system, presents an innovative solution for safeguarding the privacy of financial transaction data [29]. Its structure comprises six layers: the application layer, smart contract layer, incentive layer, consensus layer, network layer, and data layer, as depicted in Figure 4.

In the blockchain architecture diagram depicted in Figure 4, the application layer involves programmable currency and programmable finance applications provided by the Ethereum platform through smart contracts. The contract layer utilizes a scripting language unique to the Ethereum blockchain, facilitating interaction with the upper-level application and lower-level data layers. The incentive layer comprises mechanisms involving cryptographic currencies, incentivizing participants to operate and maintain the blockchain network. The consensus layer ensures a more efficient consensus on data within blocks in a fully distributed environment, employing various algorithms as indicated in Table 1. The network layer is primarily responsible for node connectivity, transaction broadcasting, and block broadcasting. The data layer predominantly employs distributed data structures to store and manage information within the blockchain [30,31].

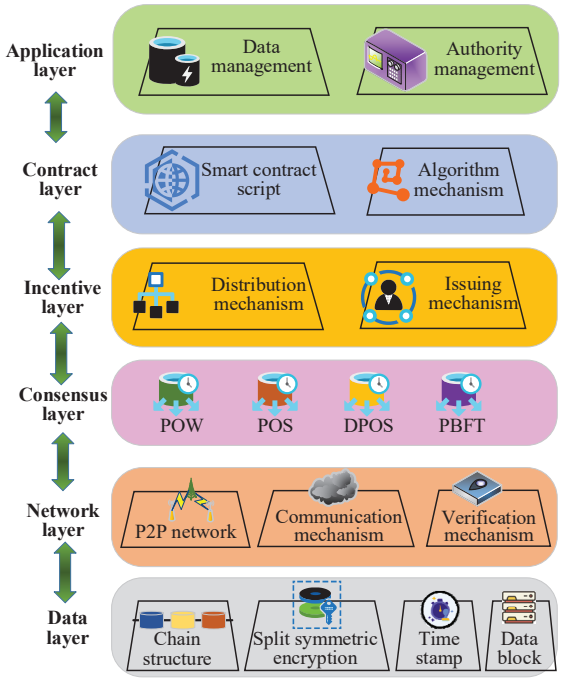


Figure 4. Sketch map of the geographical location of the study area.

Table 1. Summary of consensus algorithms in blockchain.

Algorithm	Type	Features
Proof of Work (POW)	POW	- Ensures data immutability - Requires significant computational power, high energy consumption - Adopted by blockchains like Bitcoin
Proof of Stake (POS)	POS	- Determines stake and block generation based on the quantity of tokens held - Higher energy efficiency, more environmentally friendly
Delegated Proof of Stake (DPOS)	DPOS	- Involves the election of a group of representatives for block validation and generation - Faster transaction confirmation times - Used by blockchains like EOS
Practical Byzantine Fault Tolerance (PBFT)	PBFT	- Provides a solution to the Byzantine fault tolerance problem in distributed systems - Suitable for smaller-scale node networks

As presented in Table 1, the Proof of Work (PoW) consensus algorithm is a prevalent method in the consensus layer of blockchain. It identifies the block owner through the calculation of the difficulty value, ensuring the security of user data in the financial domain and preventing malicious activities and data tampering [32]. In the event of information conflicts, PoW places trust in the chain branch with the highest number of blocks. Consequently, a malicious attacker must ensure that the length of the attack chain they send surpasses that of the main chain to acquire the authority to tamper with the entire blockchain [33]. The probability P_n can be expressed as Equation (10):

$$P_n = \begin{cases} 1, q_F \geq p_T \\ (q_F/p_T)^n, q_F < p_T \end{cases} \quad (10)$$

In Equation (10), p_T represents the probability that the next block belongs to an honest node, q_F represents the probability that the next block belongs to a malicious node, and P_n represents the probability of a malicious attacker tampering with n blocks. Expanding on this, assuming that an honest node can generate a block every average expected time, the potential progress of malicious nodes should follow a Poisson distribution, with an expected value λ. As shown in Equation (11):

$$\lambda = nq_F / p_T \tag{11}$$

Thus, when the number of blocks published by malicious nodes exceeds the number of blocks published by honest nodes, the malicious nodes have successfully conducted an attack. The probability P of this situation occurring should be determined by multiplying the probability density of the Poisson distribution of the number k of blocks already published by malicious nodes with the probability that malicious nodes can still complete the attack at that point. The specific calculation results are given by Equation (12):

$$P = \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} \begin{cases} (q_F / p_T)^{n-k} & k \leq n \\ 1 & k > n \end{cases} \tag{12}$$

Further simplify the occurrence probability P, and the result is as shown in Equation (13):

$$P = 1 - \sum_{k=0}^{n} \frac{\lambda^k e^{-\lambda}}{k!} \left[1 - (q_F / p_T)^{n-k} \right] \tag{13}$$

Through computational analysis, when the computational power of malicious nodes is below 50%, the probability of a successful attack decreases with an increase in n. In the context of specific applications for securing enterprise intelligent financial data, n is often set to 6. This implies that information from a new block is considered safe and valid only after undergoing six block generations following its creation.

Therefore, when applying blockchain technology to the privacy protection of user data in the financial domain, this article selects the PoW consensus algorithm to offer a more comprehensive and reliable safeguard for the privacy and security of financial data.

3.3. Construction and Analysis of a Financial Anti-Fraud Model Integrating Blockchain with GNN

In the domain of fraud detection within the financial sector, this article introduces the GAN from GNNs to effectively identify fraudulent information in user transactions. The Node Attention Network (NAN) is combined with the Semantic Attention Network, designing a Dual-Head Attention Network module. Additionally, to better handle non-linear features and the complex interactive relationships within data, this study integrates the GBDT. This integration results in the GBDT-DGAN network, which is utilized to detect fraudulent activities in user transactions.

Furthermore, blockchain technology is introduced to ensure the effective protection of private information during user transactions. In the end, a financial anti-fraud model based on the integration of blockchain with the GBDT-DGAN algorithm is constructed, as depicted in Figure 5.

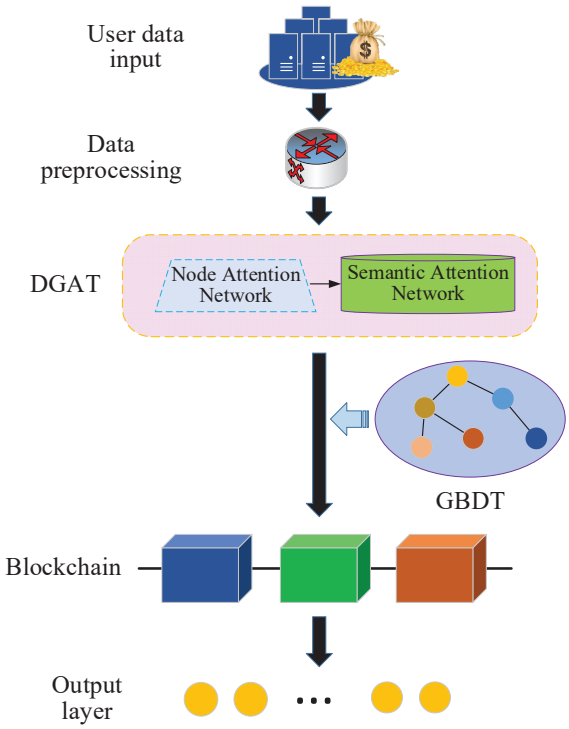


Figure 5. Framework of the financial anti-fraud model based on the blockchain-integrated GBDT-DGAN algorithm.

As illustrated in Figure 5, upon inputting user transaction data, the initial step involves preprocessing financial transaction data. This encompasses data cleaning, feature extraction, and transformation, converting transaction data into a graph representation with nodes and edges. Subsequently, the Dual-Head Attention Network module, combining the NAN and the Semantic Attention Network, is employed to handle transaction relationships within the data. The data processed through DGAN is then fused with the GBDT, specifically utilizing the GBDT-DGAN network to identify data information features of user transactions in the network. Finally, the model output data is integrated with blockchain technology, recording it on the blockchain to protect privacy information and ensure data integrity. In the blockchain network, zero-knowledge proof technology is utilized to construct trusted zero-knowledge proofs [34]. This enables the network to publish keywords required by users, regulatory agencies, and financial institutions for querying. Smart contracts verify whether the user transaction data provided in the blockchain meets anti-fraud requirements. Ultimately, the transaction data receives a label indicating fraud or security after model prediction.

In this framework model, the constructed GBDT-DGAT algorithm combines DGAT and GBDT and trains a GNN g_θ on graph G with node feature X through minimization of $L_{DGAT}(g_\theta(G, X), y, \phi, K)$ and gradient descent. It also optimizes the node feature X' and minimizes the loss function $L_{GBDT}(f^i(X'), y)$ through a dual-attention mechanism, establishes a GBDT model $f^1(X')$ for f decision trees, and updates the node feature X to X'' using the GBDT model $f^1(X')$.

In the second iteration of the model algorithm, the initial input node feature X and the updated feature X' are used as inputs to the DGAT algorithm. The updated node feature is represented as X''_{new}, and the difference between the node feature (as shown in Equation (14)) and the new node feature X''_{new} optimized using the dual-channel attention mechanism is used as the label for constructing the next decision tree, as shown in Equation (15):

$$X' = f^1(X) \tag{14}$$

$$X''_{new} = X' - \eta \frac{L_{DGAT}(g_\theta(G, X), y, \phi, K)}{\partial X'} \tag{15}$$

GBDT is used as the gradient descent update of node feature X by DGAT. The aggregation result of predicted values is shown in Equation (16):

$$f(X) = f^1(X) + f^2(X) \tag{16}$$

f^2 is the predictive performance of DGAT on this basis of f^1.

4. Experiments

4.1. Experimental Environments and Evaluation

To validate the performance of the financial anti-fraud model based on the blockchain-integrated GBDT-DGAN algorithm constructed in this article, the data source utilized was the YelpChi dataset "https://www.datafountain.cn/datasets (accessed on 25 March 2021)". The data in this dataset were desensitized and preprocessed through normalization, data cleaning, and transformation, resulting in 13,481 data points, which were split into a training set and a testing set in a 7:3 ratio. In this experiment, assuming that certain users frequently post a large number of false comments within a specific time window, this behavior may be identified as abnormal behavior that does not conform to normal patterns. The model can capture these abnormal features, such as the frequency of comments and the similarity of content. For example, among merchants in the dataset, certain behaviors that frequently engage in high-value transactions during specific time periods are considered abnormal trading patterns by the model, deviating from normal trading behavior. This may indicate the existence of potential fraud risks.

For example, user A in the dataset posted 100 comments within 24 h, while the average number of comments for most users is 10. This has been identified as abnormal behavior and may involve false comments. User B posted 50 comments within an hour, which were clearly different from normal user behavior and marked as abnormal. There are also examples of abnormal transaction patterns, such as merchant X conducting 10 high-value transactions in a short period of time at night, while other merchants have more stable transaction patterns. This abnormal behavior may imply potential fraudulent activities. Merchant Y frequently engages in micro-transactions during non-peak hours, which is inconsistent with the normal trading mode of the merchant and may be an indicator of abnormal behavior. By identifying these abnormal behaviors through models, financial institutions can further review and take appropriate measures to reduce potential fraud risks. Below are some cases of normal and abnormal modes, as shown in Tables 2–4. These tables provide detailed information about comment frequency, content similarity, and trading behavior, including various indicators of users/user pairs and whether they are identified as abnormal.

Table 2. Comment frequency distribution.

Users	Number of Comments (within Hours)	Average Comment Length	Last Activity Time	Abnormal Identification
User A	100	50	15 January 2022	Yes
User B	50	40	14 January 2022	No
User C	10	60	15 January 2022	No
User D	80	45	15 January 2022	Yes
User E	120	55	14 January 2022	Yes
User F	30	48	15 January 2022	No
User G	60	52	14 January 2022	No
User H	90	42	15 January 2022	Yes
User I	25	58	14 January 2022	No
User J	110	49	15 January 2022	Yes

Table 3. Content similarity distribution.

User Pairs	Content Similarity	Abnormal Identification
(User 1, User 2)	0.8	Yes
(User 1, User 3)	0.5	No
(User 2, User 3)	0.9	Yes
(User 4, User 5)	0.7	No
(User 5, User 6)	0.85	Yes
(User 7, User 8)	0.6	No
(User 8, User 9)	0.75	Yes
(User 10, User 11)	0.88	Yes

Table 4. Analysis of trading behavior.

Time	Transaction Volume of Merchant X	Transaction Volume of Merchant Y	Abnormal Identification
00:00–01:00	1000	500	Yes
01:00–02:00	500	2000	No
02:00–03:00	800	300	Yes
03:00–04:00	1200	400	Yes
04:00–05:00	600	600	No

In the experiments, the Hyperledger Fabric consortium blockchain platform was employed. The CPU model running on the computer was CORE-i7-4720HQ-2.6GHz. Matrix operations were performed using the open-source tools Numpy and Pandas. Numpy is an open-source matrix-processing library, and Pandas provides excellent support for data cleaning and preprocessing in data analysis. The chaincode (smart contract) was implemented in the Go language. The Pandas library in Python was instrumental in data cleaning and preprocessing. Simultaneously, the TensorFlow platform was introduced to construct the GBDT-DGAN network, utilizing various modules provided by Python. Specific hyperparameter settings were as follows: The L2 regularization parameter was set to 0.0005, and the dropout for each layer was set at 0.6, with a total of 100 iterations. The optimizer used was Adam, which applied the stochastic gradient descent algorithm to optimize the loss function with an initial learning rate of 0.001.

To assess the performance of the model proposed in this article, the model algorithm was compared with GCN, BiLSTM [35], CNN [36], and the model algorithm proposed by Usman et al. [17]. (2023) from related literature. An evaluation was conducted based on convergence, precision, recall, F1 score, and data transfer security metrics.

4.2. Analysis of Recognition Accuracy Results for Different Algorithms

The recognition performance of the model proposed in this article is compared with GCN, BiLSTM, CNN, and the model algorithm proposed by Usman et al. (2023) [17] from related literature. The convergence situations are illustrated in Figure 6. Further comparisons are made for precision, recall, and F1 score, and the results are presented in Figures 7–9.

As depicted in Figure 6, the analysis of loss values for each algorithm reveals that the proposed model algorithm in this article attains the minimum loss value, stabilizing around 0.013 by the 25th iteration. In contrast, the final loss functions of other algorithms all exceed 0.054. Hence, the financial anti-fraud model based on the blockchain-integrated GBDT-DGAN algorithm proposed in this article demonstrates superior convergence, characterized by lower loss values.

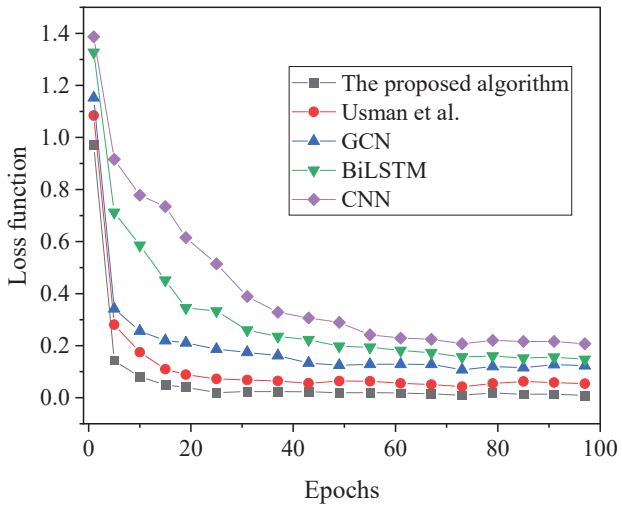


Figure 6. Convergence performance results [17].

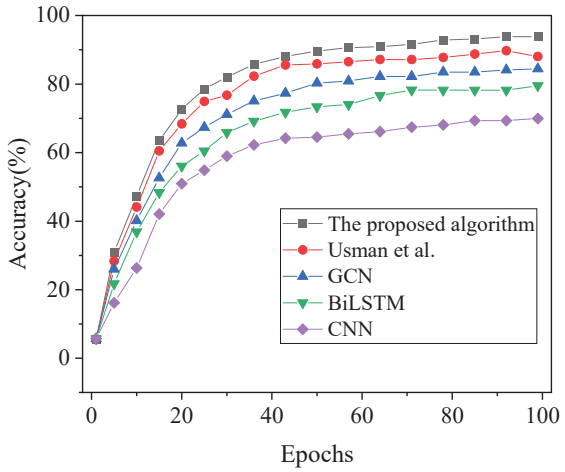


Figure 7. Accuracy results of financial fraud detection under different algorithms with varying iteration cycles [17].

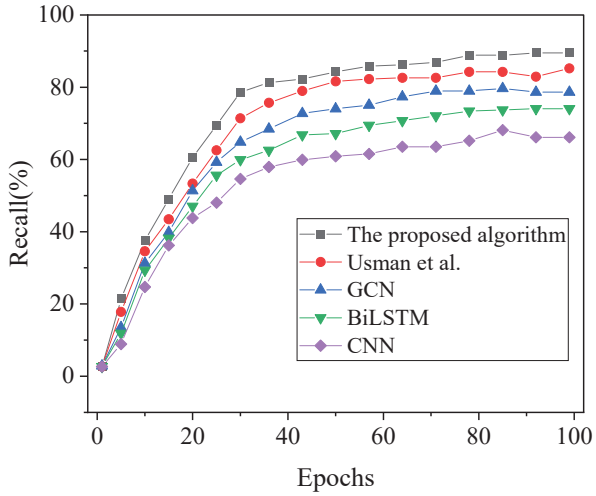


Figure 8. Recall results of financial fraud detection under different algorithms with varying iteration cycles [17].

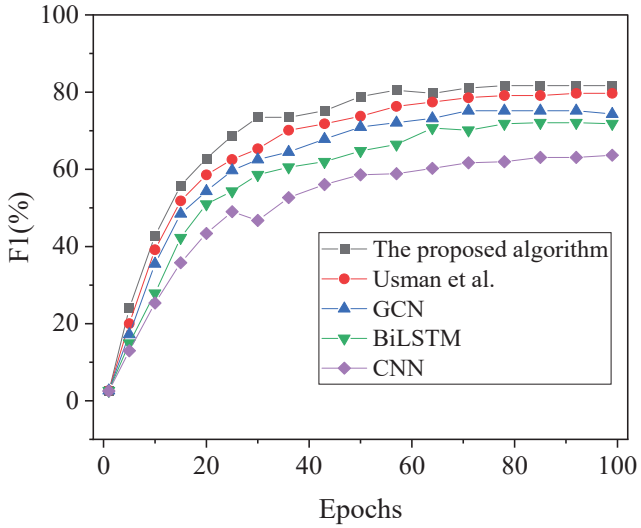


Figure 9. F1 score results of financial fraud detection under different algorithms with varying iteration cycles [17].

As illustrated in Figures 7–9, it is evident that the accuracy, recall, and F1 scores of each algorithm exhibit an initial increase followed by stabilization with the variation in iteration cycles. In the comparative analysis involving model algorithms proposed by GCN, BiLSTM, CNN, and the algorithm presented by Usman et al. (2023) [17], the accuracy of the model algorithm developed in this article reaches 93.82%. The classification recognition performance of each algorithm, ranked from highest to lowest, is as follows: model algorithm in this article > Usman et al. (2023) [17] algorithm > GCN > BiLSTM > CNN. Notably, in comparison to other algorithms, the model algorithm in this article achieves a minimum of a 5.76% increase in accuracy. Furthermore, consistent trends in recall and F1 values are observed across the algorithms, with the recall and F1 values of the model algorithm in this article reaching 89.5% and 81.66%, respectively. Consequently, the financial anti-fraud model based on the blockchain-integrated GBDT-DGAN algorithm, developed in this article, demonstrates superior predictive accuracy for identifying fraudulent data in user transactions.

Overall, the financial anti-fraud model proposed here based on the blockchain-integrated GBDT-DGAN algorithm has achieved significant advantages in multiple indicators such as loss value, accuracy, recall rate, and F1 score. This model not only performs excellently in convergence but also surpasses other comparative algorithms in accuracy and overall performance. This proves that the proposed algorithm has excellent predictive performance in accurately identifying fraudulent data in user transactions.

4.3. Security Performance Analysis of the Model under Different Algorithms

In this section, a comprehensive analysis of the security performance of the proposed model algorithm is conducted, comparing it with GCN, BiLSTM, CNN, and the algorithm presented by Usman et al. (2023) [17]. The evaluation focuses on data transmission security, as illustrated in Figure 10.

The examination of network data security transmission performance under varying data volumes reveals compelling insights. As depicted in Figure 10, the packet loss rate of the model proposed in this article demonstrates consistent performance, maintaining below 7% even as the volume of transmitted data increases. In contrast to alternative algorithms, the proposed model exhibits a lower packet loss rate, while other algorithms experience rates exceeding 17.3%. In summary, considering different data volumes, the financial anti-fraud model based on the blockchain-integrated GBDT-DGAN algorithm, as constructed in this article, showcases a significantly reduced packet loss rate, establishing its excellence in network data security transmission performance. This performance advantage is crucial for

the secure transmission of financial transaction data. While maintaining a low packet loss rate, this model can effectively ensure the integrity and confidentiality of user transaction information. This is essential for financial institutions, especially when dealing with large amounts of transaction data; ensuring the reliability of network transmission is crucial for preventing fraudulent behavior and protecting user privacy. Therefore, the proposed financial anti-fraud model not only has a breakthrough in fraud detection performance but also performs excellently in network data security transmission performance. This makes the model more widely feasible and practical in practical financial applications, providing a comprehensive, efficient, and secure anti-fraud solution for the financial field.

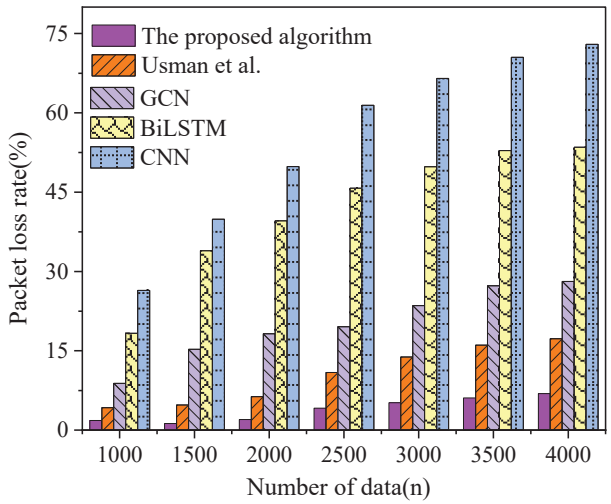


Figure 10. Security results of each algorithm under different data amounts [17].

5. Discussion

The target audience is mainly stakeholders in the financial sector, including financial institutions and regulatory agencies, as well as individuals and enterprises related to financial transactions. For financial institutions, it is a significant advancement that helps improve the security and reliability of financial transactions and reduces potential fraud risks. But data quality and quantity may be a potential challenge, as models rely heavily on high-quality and sufficient data. In addition, real-time and latency issues in the financial environment may affect the application of models, especially for trading environments that require rapid responses. The interpretability of the model may also become a key issue, and financial institutions may need to have a clear understanding of the basis for model judgments to meet regulatory requirements or customer needs. For individuals and businesses involved in financial transactions, this model will also provide them with a safer trading environment to enhance their confidence in financial transactions and reduce losses caused by fraudulent behavior.

Therefore, although the financial anti-fraud model has achieved encouraging results in experiments, some limitations still need to be addressed. Firstly, the model may have potential biases and errors, which can arise from biases in the data collection process, incomplete feature selection, or excessive reliance on certain types of fraudulent behavior. To address these issues, this article proposes a more diverse dataset collection and feature engineering optimization as a response. Secondly, data quality issues may affect the performance and robustness of the model. To address this issue, this article suggests conducting more thorough data cleaning and preprocessing, while improving the accuracy and consistency of data annotation. Finally, the model may perform well in specific financial environments, but its generalization ability is limited in other environments. Hence, it is recommended to adjust and optimize the model parameters to improve the adaptability of the model in different environments and conduct cross-environment validation to evaluate its universality.

6. Conclusions

This article introduces an advanced financial anti-fraud model, namely the GBDT–Dual-channel Graph Attention Network (GBDT-DGAN), which utilizes blockchain technology to enhance privacy protection. The structure of the GAN constructs a Dual-Head Attention Network module by integrating a Node Attention Network and a Semantic Attention Network. This module enables the comprehensive analysis of the complex relationships in user transaction data and skillfully deals with nonlinear characteristics and complex data interaction relationships. After verification, the model achieves a significant accuracy of 93.82%, which is at least 5.76% higher than baseline algorithms such as CNN. The recall rate and F1 score further demonstrate the effectiveness of the model, reaching 89.5% and 81.66%, respectively. In addition, the model performs excellently in ensuring network data transmission security, maintaining a packet loss rate of less than 7%. The experimental evaluation consolidates the superiority of the proposed model in terms of accuracy, emphasizing its potential as a powerful solution for financial transaction fraud detection. The introduction of the GBDT-DGAN algorithm, combined with blockchain technology, not only improves accuracy but also significantly enhances user privacy.

Although this article has achieved significant results, there are still some limitations. For example, this article uses the YelpChi dataset for experimentation. However, this dataset still has its limitations and may not fully cover all complex transaction patterns in the financial field. Future research can consider more extensive financial transaction data to improve the generalization of the model. Although feature engineering and data preprocessing are conducted in the article, there are still limitations that may not cover all potential patterns. Future research can further explore more refined feature engineering methods to improve the model's ability to identify different fraudulent behaviors.

Author Contributions: Conceptualization, S.W.; Data curation, S.W.; Formal analysis, S.W.; Funding acquisition, S.L.; Investigation, S.W.; Methodology, S.W.; Project administration, S.W. and S.L.; Resources, S.W.; Software, S.W.; Supervision, S.L.; Validation, S.W.; Visualization, S.W.; Writing—original draft, S.W.; Writing—review and editing, S.W. and S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) (2021RIS-001 (1345370811)).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GAN	Generative Adversarial Network
GBDT	Gradient-Boosting Decision Trees
GBDT-DGAN	Gradient-Boosting Decision Trees–Dual-channel Graph Attention Network
NAN	Node Attention Network
BiLSTM	Bidirectional Long Short-Term Memory
GNN	Graph Neural Network
GCN	Graph Convolutional Network
CNN	Convolutional Neural Network
GAT	Graph Attention Network

PoW	Proof of Work
POS	Proof of Stake
DPOS	Delegated Proof of Stake
PBFT	Practical Byzantine Fault Tolerance

References

1. Ali, A.; Razak, S.A.; Othman, S.H.; Eisa, T.A.E.; Al-Dhaqm, A.; Nasser, M.; Elhassan, T.; Elshafie, H.; Saif, A. Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review. *Appl. Sci.* **2022**, *12*, 9637. [CrossRef]
2. Ragazou, K.; Passas, I.; Garefalakis, A. It Is Time for Anti-Bribery: Financial Institutions Set the New Strategic "Roadmap" to Mitigate Illicit Practices and Corruption in the Market. *Adm. Sci.* **2022**, *12*, 166. [CrossRef]
3. Hendri, N.; Sari, S.U. Sistematic Literature Review: The Strategy for Preventing Government Financial Report Fraud. *JAK (J. Akunt.) Kaji. Ilm. Akunt.* **2023**, *10*, 323–336. [CrossRef]
4. Li, J.; Yang, D. Research on Financial Fraud Detection Models Integrating Multiple Relational Graphs. *Systems* **2023**, *11*, 539. [CrossRef]
5. Ma, X.; Wu, J.; Xue, S.; Yang, J.; Zhou, C.; Sheng, Q.Z.; Xiong, H.; Akoglu, L. A comprehensive survey on graph anomaly detection with deep learning. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 12012–12038. [CrossRef]
6. Ashtiani, M.N.; Raahemi, B. Intelligent fraud detection in financial statements using machine learning and data mining: A systematic literature review. *IEEE Access* **2021**, *10*, 72504–72525. [CrossRef]
7. Li, Y.; Zhao, J.; Lv, Z.; Li, J. Medical image fusion method by deep learning. *Int. J. Cogn. Comput. Eng.* **2021**, *2*, 21–29. [CrossRef]
8. Li, J.; Chang, Y.; Wang, Y.; Zhu, X. Tracking down financial statement fraud by analyzing the supplier-customer relationship network. *Comput. Ind. Eng.* **2023**, *178*, 109118. [CrossRef]
9. Huo, H.; Guo, J.; Yang, X.; Lu, X.; Wu, X.; Li, Z.; Li, M.; Ren, J. An Accelerated Method for Protecting Data Privacy in Financial Scenarios Based on Linear Operation. *Appl. Sci.* **2023**, *13*, 1764. [CrossRef]
10. Liu, H.; Jiang, R. A Causal Graph-Based Approach for APT Predictive Analytics. *Electronics* **2023**, *12*, 1849. [CrossRef]
11. Fang, W.; Li, X.; Zhou, P.; Yan, J.; Jiang, D.; Zhou, T. Deep learning anti-fraud model for internet loan: Where we are going. *IEEE Access* **2021**, *9*, 9777–9784. [CrossRef]
12. Alarfaj, F.K.; Malik, I.; Khan, H.U.; Almusallam, N.; Ramzan, M.; Ahmed, M. Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. *IEEE Access* **2022**, *10*, 39700–39715. [CrossRef]
13. Cheng, D.; Xiang, S.; Shang, C.; Zhang, Y.; Yang, F.; Zhang, L. Spatio-temporal attention-based neural network for credit card fraud detection. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 362–369. [CrossRef]
14. Lei, X.; Mohamad, U.H.; Sarlan, A.; Shutaywi, M.; Daradkeh, Y.I.; Mohammed, H.O. Development of an intelligent information system for financial analysis depend on supervised machine learning algorithms. *Inf. Process. Manag.* **2022**, *59*, 103036. [CrossRef]
15. Xu, L.; Wang, J.; Xu, D.; Xu, L. Integrating Individual Factors to Construct Recognition Models of Consumer Fraud Victimization. *Int. J. Environ. Res. Public Health* **2022**, *19*, 461. [CrossRef] [PubMed]
16. Verykios, V.S.; Stavropoulos, E.C.; Zorkadis, V.; Katsikatsos, G.; Sakkopoulos, E. Sensitive data hiding in financial anti-fraud process. *Int. J. Electron. Gov.* **2022**, *14*, 7–27. [CrossRef]
17. Usman, A.; Naveed, N.; Munawar, S. Intelligent Anti-Money Laundering Fraud Control Using Graph-Based Machine Learning Model for the Financial Domain. *J. Cases Inf. Technol.* **2023**, *25*, 1–20. [CrossRef]
18. Cheng, D.; Yang, F.; Xiang, S.; Liu, J. Financial time series forecasting with multi-modality graph neural network. *Pattern Recognit.* **2021**, *121*, 108218. [CrossRef]
19. Zhang, Q.; Zhang, Y.; Yao, X.; Li, S.; Zhang, C.; Liu, P. A Dynamic Attributes-driven Graph Attention Network Modeling on Behavioral Finance for Stock Prediction. *ACM Trans. Knowl. Discov. Data* **2023**, *18*, 1–29. [CrossRef]
20. Xu, C.; Huang, H.; Ying, X.; Gao, J.; Li, Z.; Zhang, P.; Xiao, J.; Zhang, J.; Luo, J. HGNN: Hierarchical graph neural network for predicting the classification of price-limit-hitting stocks. *Inf. Sci.* **2022**, *607*, 783–798. [CrossRef]
21. Lazcano, A.; Herrera, P.J.; Monge, M. A Combined Model Based on Recurrent Neural Networks and Graph Convolutional Networks for Financial Time Series Forecasting. *Mathematics* **2023**, *11*, 224. [CrossRef]
22. Sun, Q.; Wei, X.; Yang, X. GraphSAGE with deep reinforcement learning for financial portfolio optimization. *Expert Syst. Appl.* **2024**, *238*, 122027. [CrossRef]
23. Varlamis, I.; Michail, D.; Glykou, F.; Tsantilas, P. A survey on the use of graph convolutional networks for combating fake news. *Future Internet* **2022**, *14*, 70. [CrossRef]
24. Michail, D.; Kanakaris, N.; Varlamis, I. Detection of fake news campaigns using graph convolutional networks. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100104. [CrossRef]
25. Ghadimi, A.; Beigy, H. SGCSumm: An extractive multi-document summarization method based on pre-trained language model, submodularity, and graph convolutional neural networks. *Expert Syst. Appl.* **2023**, *215*, 119308. [CrossRef]
26. Zhou, X.; Yang, W.; Tian, X. Detecting Phishing Accounts on Ethereum Based on Transaction Records and EGAT. *Electronics* **2023**, *12*, 993. [CrossRef]
27. Hu, X.; Chen, H.; Chen, H.; Li, X.; Zhang, J.; Liu, S. Mining Mobile Network Fraudsters with Augmented Graph Neural Networks. *Entropy* **2023**, *25*, 150. [CrossRef]

28. Zhang, Z.; Luo, M.; Hu, Z.; Niu, H. Textual Emotional Tone and Financial Crisis Identification in Chinese Companies: A Multi-Source Data Analysis Based on Machine Learning. *Appl. Sci.* **2022**, *12*, 6662. [CrossRef]
29. Kumari, V.; Bala, P.K.; Chakraborty, S. An Empirical Study of User Adoption of Cryptocurrency Using Blockchain Technology: Analysing Role of Success Factors like Technology Awareness and Financial Literacy. *J. Theor. Appl. Electron. Commer. Res.* **2023**, *18*, 1580–1600. [CrossRef]
30. Cucari, N.; Lagasio, V.; Lia, G.; Torriero, C. The impact of blockchain in banking processes: The Interbank Spunta case study. *Technol. Anal. Strateg.* **2022**, *34*, 138–150. [CrossRef]
31. Metawa, N.; Alghamdi, M.I.; El-Hasnony, I.M.; Elhoseny, M. Return Rate Prediction in Blockchain Financial Products Using Deep Learning. *Sustainability* **2021**, *13*, 11901. [CrossRef]
32. Li, J.; Chen, Y.; Chen, Y.; Zhang, W.; Liu, Z. A smart energy IoT model based on the Itsuku PoW technology. *Results Eng.* **2023**, *18*, 101147. [CrossRef]
33. Ramos, S.; Pianese, F.; Leach, T.; Oliveras, E. A great disturbance in the crypto: Understanding cryptocurrency returns under attacks. *Blockchain Res. Appl.* **2021**, *2*, 100021. [CrossRef]
34. Jiang, S.; Cao, J.; Wu, H.; Chen, K.; Liu, X. Privacy-preserving and efficient data sharing for blockchain-based intelligent transportation systems. *Inf. Sci.* **2023**, *635*, 72–85. [CrossRef]
35. Wang, M.; Xie, Z.; Wen, X.; Li, J.; Zhou, K. Ethereum Smart Contract Vulnerability Detection Model Based on Triplet Loss and BiLSTM. *Electronics* **2023**, *12*, 2327. [CrossRef]
36. Cheah, P.C.Y.; Yang, Y.; Lee, B.G. Enhancing Financial Fraud Detection through Addressing Class Imbalance Using Hybrid SMOTE-GAN Techniques. *Int. J. Financ. Stud.* **2023**, *11*, 110. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

International Journal of
Financial Studies

MDPI

Article

Board Expertise Background and Firm Performance

Chiou-Yann Lee, Chun-Ru Wen and Binh Thi-Thanh-Nguyen *

Department of Accounting, Chaoyang University of Technology, 168 Jifong E. Road, Wufong District, Taichung City 41349, Taiwan; bear21fish@gmail.com (C.-R.W.)
* Correspondence: tbnguyen@cyut.edu.tw

Abstract: This study presents a novel financial performance forecasting method that combines the threshold technique with Artificial Neural Networks (ANN). It applies the threshold regression method to identify the factors within the board of directors that influence the financial performance of traditional industries in Taiwan. The findings indicate that the ANN method effectively predicts financial performance by using relevant board structure data. Furthermore, the empirical results suggest that boards with more members demonstrate increased profitability. Additionally, a more significant presence of board members with accounting expertise contributes to more consistent profits. In contrast, an increased presence of members with financial expertise has a more pronounced impact on profitability.

Keywords: artificial neural networks; multi-threshold model; financial performance

1. Introduction

Traditional industries, including sectors such as metal machinery, the chemical industry, and subsistence industries, have played pivotal roles in Taiwan's economic development, trade, and industrial evolution over the years. However, in recent years we have witnessed the emergence of economic powerhouses and the rapid progress of regional economic and trade partnerships like ASEAN, resulting in significant changes to the global economic and trade landscape. As a result, these industries are now operating within a new and highly competitive environment. In this era of rapid transformation, these industries must identify the essential factors that will enable them to maintain and enhance their positions on the international stage.

Investors seek to achieve multiple objectives through the election of a company's board of directors. They aim to enhance corporate governance, safeguard the rights and interests of shareholders, and anticipate that the board will contribute expertise to improve the decision-making prowess of the board itself. Ultimately, this is expected to enhance the company's overall performance. Consequently, whether the board of directors can positively impact corporate performance has remained a significant research focus in management and academia.

While many studies have explored the relationship between the board of directors' characteristics and company performance, a notable need exists for more in-depth articles examining the professional backgrounds of the board members and CEOs and how their leadership qualities influence financial performance. This gap in research underscores the need for a more comprehensive understanding of the intricate dynamics between leadership, professional expertise, and corporate financial outcomes.

Over the past four decades, numerous studies have identified the common attributes of effective corporate boards. Board structure has consistently taken center stage in scholarly discussions among the myriad factors explored. The primary roles of a board of directors encompass leadership and decision-making. Board members hold a prominent position where they oversee the bigger picture, conduct macro-level analysis and assessments in line with the organization's overarching vision, steer the general direction of the company,

and guide the company's proactive responses to societal shifts. The fundamental purpose of the board of directors is to govern the company, making it the backbone of the entire organization. Jensen and Zajac (2004) have contended that boards comprising highly educated members tend to be more amenable to embracing changes in a company's structure or strategy. This receptivity arises from their ability to mitigate systemic risks associated with future changes, owing to their unique strategic perspectives, nimble problem-solving abilities, and in-depth comprehension of their respective businesses. Numerous studies, including those conducted by Hambrick and Mason (1984) and Kor and Sundaramurthy (2009), have reached a consensus that top management possessing specialized knowledge, particularly a deep understanding of the business and its operational intricacies, tends to lead to more effective business operations. Educational qualifications serve as a crucial indicator of knowledge, with degrees in finance, accounting, economics, research, and law being particularly relevant due to their extensive application in corporate governance.

Recent studies provide compelling evidence for the intricate link between board composition and performance. Amini and Zhang (2022) explored the impact of "industry-specific human capital" on board composition, demonstrating a positive correlation between relevant board member experience and firm risk management effectiveness. Similarly, Aguilera et al. (2021) highlight the importance of "functional expertise" on boards, finding that firms with directors possessing financial expertise exhibit superior financial performance, particularly during periods of economic downturn. Beyond expertise, the broader background of board members also plays a crucial role. A study by Francoeur et al. (2021) examined the impact of "boardroom gender diversity" on firm performance, revealing a positive association between diverse boards and higher return on assets (ROA) and return on equity (ROE). Likewise, Luo et al. (2020) investigated the effect of "board age diversity", with results suggesting that a balanced age mix on boards can enhance firm innovation and performance, particularly in technology-intensive industries. However, the relationship between board makeup and performance remains nuanced. As Klein and Zurbrügg (2023) pointed out, the effectiveness of specific board characteristics hinges on various moderating factors, including firm size, industry context, and corporate governance practices. Moreover, boardroom dynamics, leadership, and information flow between the board and management play significant roles in translating individual expertise and diverse backgrounds into tangible performance gains.

This paper emphasizes the critical role of board composition in organizational success. Business managers are advised to recruit board members with diverse expertise in accounting and finance to enhance decision-making and financial oversight. They should carefully evaluate board size to balance diverse perspectives with efficient communication. Additionally, managers must monitor for unethical profit management practices, leveraging predictive analytics for strategic insights. Contextualizing their findings within their specific industry and organizational context is crucial for effective implementation and sustained growth.

In pursuing this investigation, this paper has harnessed a vast dataset comprising 5309 observations across 331 companies, spanning from 2000 to 2021, as its research foundation. The employed methodology centers on applying the threshold analysis technique to unveil the influence of board characteristics on business performance, focusing on identifying overarching trends regarding Return on Assets (ROA). These findings subsequently serve as the bedrock for developing an Artificial Neural Network (ANN) tailored for forecasting financial performance. Compared to existing research, this paper sets itself apart through its innovative approach, offering a novel research model within this domain.

This paper represents one of the pioneering attempts to meld the ANN model with the multi-threshold technique to predict a firm's ROA based on the attributes of its board of directors. The research outcomes reveal that a board of directors boasting a more significant number of members with accounting qualifications contributes to a consistent and stable increase in ROA. Conversely, an augmented presence of members with financial expertise yields a more pronounced yet less stable impact on ROA. Moreover, members

with legal expertise appear to exert no significant favorable influence on ROA. Furthermore, directors holding more company shares exhibit the capacity to enhance ROA, but this effect materializes only when ROA surpasses the threshold of 14.96%. Additionally, when considering the factor of simultaneously serving on the board of directors, the positive effect becomes more prominent. In essence, directors who are actively engaged in managing and participating in the company's decision-making processes exhibit a heightened ability to bolster profitability.

2. Related Literature

Fueled by market growth and increasing complexity, the ever-expanding business landscape presents firms with a critical challenge: securing essential resources for survival and success. Resource Dependence Theory (RDT) offers a powerful lens through which to examine this dynamic, highlighting how organizations rely on external resources and how this dependence shapes their actions and, ultimately, their performance. Within this framework, boards of directors emerge as critical players tasked with strategically acquiring and managing these resources.

RDT emphasizes the significance of board expertise in building bridges to critical resources. Boards with members possessing relevant industry experience, financial literacy, and functional expertise are better equipped to attract investments, partnerships, and talent (Krishnan and Santos 2008; Rynes et al. 2007). This expertise empowers them to assess opportunities, build trust with external stakeholders, and navigate complex resource dependencies effectively. By leveraging their knowledge and networks, they can unlock doors to critical resources that drive firm growth and competitiveness.

However, RDT goes beyond mere expertise, recognizing the value of board diversity in accessing a more comprehensive range of resources. The educational backgrounds, professional affiliations, and past experiences of board members bring unique perspectives and networks to the table (Boone and Lüdema 2018). This diversity fosters innovation, enhances problem-solving, and opens doors to previously untapped resources, ultimately enriching the firm's resource portfolio. By embracing diversity, boards gain access to a broader spectrum of knowledge, connections, and experiences, allowing them to identify and secure resources that may have been overlooked with a more homogenous composition.

RDT acknowledges that resource dependence is not without its nuances. While larger boards offer the potential for broader expertise and diverse perspectives, concerns about communication complexity and efficiency can arise (Adams and Ferreira 2007). Similarly, attracting highly qualified board members might involve increased compensation, necessitating a careful cost-benefit analysis. Additionally, the effectiveness of board composition hinges on factors like firm size, industry context, and internal governance practices (Hambrick 2005). Recognizing these trade-offs and moderating factors is crucial for optimizing board composition and ensuring that resource acquisition efforts translate into tangible benefits for the firm.

Emerging research continues to solidify the link between board composition and resource acquisition, consequently impacting performance. Amini and Zhang (2022) found a positive correlation between industry-specific board experience and risk management effectiveness, highlighting the role of expertise in securing critical resources. Similarly, Aguilera et al. (2021) demonstrated that firms with financially literate boards exhibited superior financial performance, suggesting access to capital as a potential benefit. These findings, alongside others exploring the positive impact of board diversity (Francoeur et al. 2021; Luo et al. 2020), showcase the increasingly recognized role of board composition in resource acquisition and firm success.

By adopting an RDT perspective, we understand how boards navigate resource dependencies and influence firm performance. Recognizing the interplay between expertise, diversity, and moderating factors is crucial for optimizing board composition and securing the resources that drive sustainable success in our ever-evolving business landscape.

In the realm of corporate governance research, the common thread is the measurement of business performance (Hakim and Liu 2021; Li and Patel 2019). This focus stems from businesses' fundamental goal, profit generation, and the paramount importance of safeguarding shareholders' rights and interests. Additionally, discussions about the impact of corporate governance extend to areas such as earnings management (Warfield et al. 1995; Klein 2002; Chang et al. 2007), earnings forecast errors (Hlel et al. 2020), and the influence of accountants issuing going concern opinions (DeFond et al. 2002; Wu et al. 2011). Amidst this, corporate governance is bifurcated into internal and external governance, with the board of directors serving as the core of internal governance. Consequently, when exploring the supervisory mechanisms of corporate governance, the board of directors' soundness, composition, and characteristics emerge as standard independent variables.

The educational background of board members imparts diverse perspectives and mindsets regarding the company's ultimate goals, which can significantly influence its performance and purpose. Educational backgrounds offer nuanced insights into an individual's values and preferences, influenced by variations in educational programs (Hambrick and Mason 1984). The educational diversity among board members is linked to the financial performance of companies, with a stronger positive relationship observed for firms with higher levels of internationalization (Carpenter 2002). Additionally, women and directors with accounting backgrounds tend to enhance compassion and reciprocity in CSR decision-making, especially in companies with larger boards (Nguyen and Huang 2020).

The board plays a pivotal role in mitigating agency problems, and the effectiveness of this role is closely linked to its composition (Fama and Jensen 1983). Typically, the board of directors comprises internal and external directors. Since 2002, Taiwan has mandated that newly established companies must have at least two independent directors on their boards. This move aims to bring impartial and objective professionals into managerial roles, thus enhancing the quality and efficiency of board decision-making. However, concerns exist about performance evaluation's supervisory function and objectivity when internal directors are involved in a company (Weisbach 1988). Additionally, unfamiliarity with the company may lead to increased communication costs for the board (Pucheta-Martínez and Gallego-Álvarez 2020).

In recent years, directors' characteristics have become a new focus in board characteristics, including gender, expertise, education level, industrial background, and experience. Researchers have shown that senior managers' specific skills can enhance business operations from the social capital perspective, and gender diversity on boards can enhance corporate value (Kor and Sundaramurthy 2009; Kim and Starks 2016). Higher levels of education, professional knowledge, generalist skills, and experience among board members are associated with improved firm performance (Jensen and Zajac 2004; Hakim and Liu 2021). Furthermore, industry experience among directors is linked to enterprise value, and experiences from upstream and downstream industries can bring benefits in terms of information and industry expertise, potentially outweighing any concerns about conflicts of interest (Dass et al. 2014). Finally, directors from the same geographical area as the company may have stronger local connections due to their understanding of the local business environment, which can significantly impact the company's value (Sun 2021).

Various factors likely moderate the impact of board expertise and background. Board size, composition (independence, diversity), and internal firm environment (culture, leadership structure) can influence how effectively their expertise is utilized (Adams and Ferreira 2007; Hambrick 2005).

Based on the reviewed literature, we propose the following hypotheses:

Larger boards offer the potential for broader expertise and diverse perspectives, which may lead to improved decision-making and firm performance. However, concerns exist about communication complexity and potentially reduced efficiency in larger boards.

Hypothesis 1. *Boards with a higher number of members demonstrate increased profitability.*

Accounting expertise equips board members to understand financial information better, assess risk, and monitor management financial decisions, potentially leading to more consistent and stable profitability.

Hypothesis 2. *A more significant presence of board members with accounting expertise contributes to more consistent profits.*

Given the financial nature of a firm's performance, financial expertise may positively influence profitability compared to other types of expertise.

Hypothesis 3. *An increased presence of members with financial expertise has a more pronounced impact on profitability than expertise in other domains.*

This study is motivated by the pressing need to comprehensively address gaps in the existing research, explore the potential insights advanced methodologies can yield, and investigate the evolving economic landscape that demands a deeper understanding of the dynamics within boardrooms. Through these efforts, it aims to contribute knowledge that can inform both academia and industry, ultimately enhancing the grasp of the critical factors shaping corporate financial outcomes.

3. Research Methodology

The ANN architecture is derived from the idea of simulating the human brain. Like humans, ANNs learn by experience, saving and using those experiences in the right situations. An ANN usually organizes neurons into layers, each responsible for a specific task. An ANN usually has three layers, an input layer, a hidden layer, and an output layer: (1) The input layer provides the network with the necessary data. The number of neurons in the input layer corresponds to the number of input parameters given to the network, and these input parameters are assumed to be in vector form. (2) The hidden layer contains hidden neurons that connect input values to output values. A neural network may have one or more hidden layers primarily responsible for processing the neurons of the input layer and delivering information to the neurons of the output layer. These neurons are suitable for classifying and identifying the relationship between input parameters and output parameters. (3) The output layer contains output neurons which transfer the output information of computations from the ANN to the user. An ANN can be built to have multiple output parameters. The problem will decide the number of neurons in the input layer, the output layer, and the number of hidden layer neurons; the input will decide the number of hidden layers. However, choosing the type and quantity of input parameters has a class effect on the quality of the network. The mathematical model of a straight-propagation ANN is presented as follows:

$$y(x) = f\left(\sum_{i=1}^{n} w_i x_i\right) \tag{1}$$

where $y(x)$ is the output value according to the variable x; f is the activation or transfer function; w_i is the link weight of neuron x_i; x_i is the input value. Figure 1 explains the nature of the operating principle of a direct propagation ANN for the network's training. Specifically, the training process often uses a back-propagation algorithm to find the derivative for each parameter in the network.

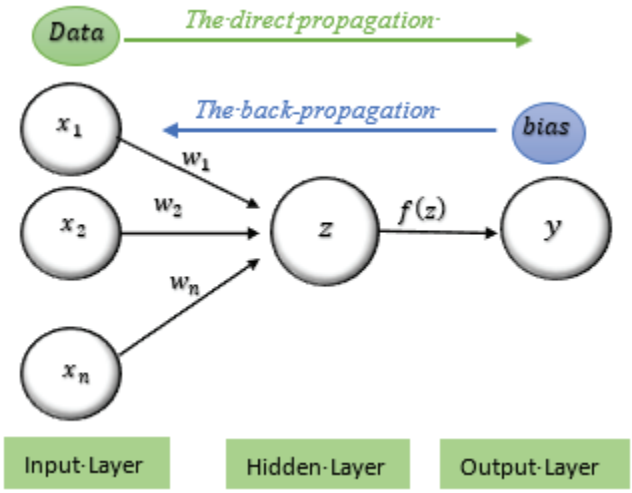


Figure 1. Operating mechanism of ANN.

The straight propagation phase consists of six steps, as follows:
The input parameter vector is input to the neurons in the input layer:

$$a^{(0)} = x \tag{2}$$

At the jth hidden layer neuron, the value of the signal received from the input layer is summed up as the weighted $(w_1, w_2 \ldots, w_n)$ sum of all the input data by adding all the products of each input data and the association weights (z) between the hidden layers and input class. b is the bias coefficient.

$$z_{inj} = b_{oj} + \sum_{i=1}^{n} x_i w_{ij} \tag{3}$$

Then, the activation function (transfer function) will be used to convert the received value into the output value.

$$z_j = f\left(z_{inj}\right) \tag{4}$$

Next, the output value at the hidden layer neuron j continues to be transmitted to the output neuron k in the same way as from the input layer to the hidden layer.

$$y_{ink} = e_{ok} + \sum_{j=1}^{p} z_j v_{jk} \tag{5}$$

Then, the transfer function is used to compute the output value of the neuron at the output layer.

$$y_k = f(y_{ink}) \tag{6}$$

The direct propagation phase ends, and the network will move to the back-propagation phase.

During the input phase, the input data includes input and actual values. Following this, for each data set, each corresponding output error is calculated. This value is called the loss function (Cost Function—J).

$$J = t_k - y_k \tag{7}$$

From the cost function just found, we calculate the derivative of this function according to the weight $(v_1, v_2 \ldots, v_n)$ between the hidden layer—the output layer, and the weight between the input layer—the hidden layer.

$$\Delta w_{ij} = \frac{\partial J}{\partial w_{ij}} \tag{8}$$

$$\Delta v_{jk} = \frac{\partial J}{\partial v_{jk}} \tag{9}$$

Next, the association weight value between the hidden layer and the output layer and the link weight value between the input layer and the hidden layer are adjusted simultaneously.

$$w_{ij}(new) = w_{ij}(old) + \alpha \Delta w_{ij} \tag{10}$$

$$v_{jk}(new) = v_{jk}(old) + \alpha \Delta v_{jk} \tag{11}$$

While ANN has a well-established presence in financial modeling, our study's originality lies in its unique combination of an ANN and multi-threshold analysis within the specific domain of corporate governance and financial performance prediction.

4. The Data and Empirical Results

4.1. Data

This study utilizes relevant data comprising 5309 observations of the boards of directors from 331 companies spanning the period from 2000 to 2021 as its research sample. These companies belong to Taiwan's traditional industries and are listed on the Taiwan Stock Exchange. The data are sourced from the Taiwan Economic Journal (TEJ). After excluding non-available values, 5309 observations are used to (i) detect significant threshold values for financial performance, (ii) present the factors influencing financial performance, and (iii) construct an Artificial Neural Network (ANN) model for predicting financial performance.

The choice of Taiwan and traditional industries as the study's focus is motivated by the unique economic context, the historical significance of these industries, the diverse challenges they encounter, and the potential policy implications. This selection enables a comprehensive and context-rich exploration of the relationship between board characteristics and financial performance in an evolving economic landscape.

This study delves into the connection between board member characteristics and financial performance, explicitly emphasizing individual education and majors (including accounting, finance, and law) as variables for measuring directors' strengths. It also considers factors such as the management level's shareholding ratio, the size of the board of directors, and the number of external directors and supervisors to predict the financial performance of Taiwan's traditional industries using an ANN model.

Table 1 summarizes the estimation and definition of variables in the empirical model, including financial performance (ROA), the professional background of the board directors (ACC, FIN, LAW), board of directors and managers (BDP, MNP1, MNP2, BSZ, IBD), and control variables (SIZE, DBT, NPT).

Table 1. Variable definition.

Variable	Definition
	Financial Performance
ROA_{it}	The return on the asset after tax, before interest, and depreciation.
	Board and Managers (BSM)
ACC_{it}	Percentage of accounting professional background of board directors, supervisors, and managers.
FIN_{it}	Percentage of finance professional background of board directors, supervisors, and managers.

Table 1. *Cont.*

Variable	Definition
LAW_{it}	Percentage of law professional background of board directors, supervisors, and managers.
BDP_{it}	A measure of the board's power. The shareholding ratio of board directors and supervisors is calculated as the total number of shares held by the board and supervisors divided by the total number of shares.
$MNP1_{it}$	A measure of the manager's power. The proportion of shares held by the company's managers in the company's announcement to the company's total issued shares. Formula: number of shares held by managers ÷ total number of shares × 100
$MNP2_{it}$	A measure of the manager's power. The proportion of shares held by all managers of the company to the total issued shares of the company refers to the shares held by managers plus the number of shares held by natural person directors who also serve as managers. Formula: number of shares held by managers (including concurrent directors and supervisors) ÷ total number of shares × 100
BSZ_{it}	The natural log of the number of directors in the board of directors.
IBD_{it}	The natural log of the number of independent directors on the board of directors.
	Control Variables (CTL)
$SIZE_{it}$	Total assets are the natural log of total assets at the end of the period.
DBT_{it}	The debt ratio is calculated as the total liabilities of total assets
NPT_{it}	Net profit rate before tax and depreciation.

Table 2 provides an overview of the main statistical parameters of the data used in the study. The mean and median values of ROA are 6.52 and 6.08, respectively, indicating that ROA is reasonably consistent across enterprises, except for instances with a maximum value of 63.3 and a minimum value of −53.12. The mean of accounting expertise (ACC) among the board of directors is 7%, higher than that of finance (FIN) and law (LAW) at 3%. There is a noticeable variation in board power (BDP) among companies, with a standard deviation of 14.84%. The mean values of MNP1 and MNP2 are 1.14 and 4.93, respectively, indicating a relatively high proportion of directors concurrently serving as board members.

Table 2. Statistics Summary.

Variables	Mean	Median	Maximum	Minimum	Std. Dev.	Obs.
ROA	6.52	6.08	63.30	−53.12	6.74	5309
ACC	0.07	0.00	0.71	0.00	0.10	5309
FIN	0.03	0.00	0.43	0.00	0.06	5309
LAW	0.03	0.00	0.44	0.00	0.06	5309
BDP	24.83	21.28	100.00	0.00	14.84	5309
MNP1	1.14	0.17	21.64	0.00	2.38	5309
MNP2	4.93	2.05	58.32	0.00	7.27	5309
BSZ	0.86	0.85	1.32	0.30	0.14	5309
IBD	0.19	0.00	0.60	0.00	0.20	5309
SIZE	6.75	6.69	8.83	4.77	0.62	5309
DBT	43.88	45.03	101.97	0.69	16.80	5309
NPT	10.65	10.46	619.96	−2989.88	49.72	5309

Table 3 analyzes the correlations between variables. The relationship between ROA and ACC exhibits a notably positive association, while the relationship between ROA and FIN shows the opposite direction. The association between ROA and LAW is not as significant. Multicollinearity, a situation where two or more predictors are highly linearly

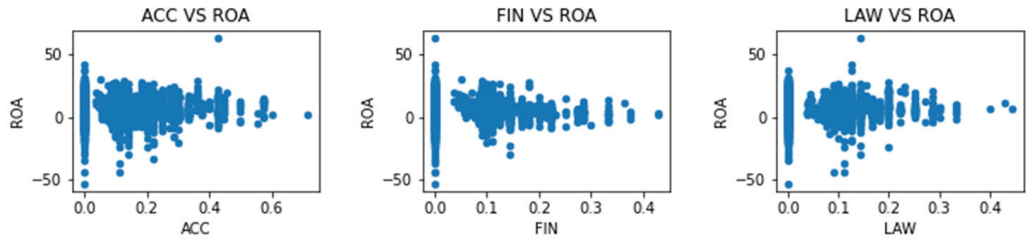
related, is assessed in this study. All absolute correlation coefficients are less than 0.6, indicating the absence of multicollinearity.

Table 3. Correlation Analysis.

Prob.	ROA		ACC		FIN		LAW		BDP		MNP1		MNP2		BSZ		IBD		SIZE		DBT	
ACC	0.07	***	1																			
FIN	−0.09	***	0.22	***	1																	
LAW	0.01		0.02		0.00		1															
BDP	0.15	***	−0.05	***	0.01		0.05	***	1													
MNP1	0.08	***	0.06	***	−0.06	***	0.03	**	−0.03	**	1											
MNP2	0.10	***	−0.03	**	−0.12	***	−0.10	***	0.18	***	0.36	***	1									
BSZ	0.05	***	0.09	***	0.09	***	0.03	**	0.03	*	−0.17	***	−0.24	***	1							
IBD	0.14	***	0.52	***	0.11	***	0.04	***	0.00		0.04	***	−0.02		0.26	***	1					
SIZE	0.10	***	0.02		0.11	***	0.11	***	−0.14	***	−0.14	***	−0.28	***	0.46	***	0.07	***	1			
DBT	−0.22	***	−0.08	***	0.03	**	−0.05	***	−0.08	***	−0.01		−0.03	**	−0.02		−0.05	***	0.16	***	1	
NPT	0.25	***	0.04	***	0.00		0.01		0.02	*	0.02		0.02		0.02		0.03	**	0.09	***	−0.05	***

Note: ***, **, and * represent the 1%, 5%, and 10% significant levels, respectively.

This study primarily focuses on the impact of board member professionalism on ROA. Figure 2 presents ROA, ACC, FIN, and LAW scatter plots.



(**a**) (**b**) (**c**)

Figure 2. Plot Scatter of Relationship between ROA and Board's Background. (**a**) ACC vs. ROA, (**b**) FIN vs. ROA, (**c**) LAW vs. ROA.

4.2. Empirical Results

4.2.1. Testing for the Threshold Model

This study uses a threshold regression model to investigate the influence of the board of directors' professional background on corporate profitability. According to Hansen (1999, 2000), the threshold regression model is built based on the data pool set. The regression is as follows:

$$ROA_{i,t} = \begin{cases} C + \delta_1 ROA_{i,t} \sum_{j=0}^{m} \Psi_j(\Phi_{it};\tau) + \beta_1 ACC_{i,t} + \beta_2 FIN_{i,t} + \beta_3 LAW_{i,t} + \beta_4 BDP_{i,t} + \beta_5 MNP1_{i,t} \\ + \beta_6 MNP2_{i,t} + \beta_7 BSZ_{i,t} + \beta_8 IBD_{i,t} + \beta_9 SIZE_{i,t} + \beta_{10} DBT_{i,t} + \beta_{11} NPT_{i,t} + \varepsilon_{i,t} \end{cases} \quad (12)$$

This study uses the OLS method to estimate the threshold regression model threshold. It uses the Bootstrap method to simulate the LRT test with asymptotic distribution to calculate the F-statistics and *p*-values of the variables to test for the existence of thresholds of the model. Each test is performed using a bootstrap of 1000 iterations. Table 4 presents the results of testing for threshold values in the Return on Assets (ROA) variable using the Bai–Perron methodology. This analysis aims to identify critical points in ROA where significant changes or shifts in a company's financial performance occur. These threshold values help to distinguish different regimes or states in the behavior of the ROA. Identifying thresholds in financial data can have important implications for understanding the dynamics of a company's financial performance, making it a valuable aspect of statistical analysis.

Table 4. Testing for the Threshold of Return on Assets (ROA).

Thresholds	τ_1	τ_2	τ_3
	14.69 ***	6.52 ***	−1.54 ***
	[29.68]	[28.40]	[26.13]

Note: The threshold variable is ROA. The maximum threshold applied for detection is 5. Statistics in [.] are critical values based on Bai and Perron (2003), which show all τ values of ROA are significant at the 1% level. *** indicates statistical significance at the 1% level.

The results analyzed above show that the impact under these thresholds exists in the sample data set under consideration. The estimated values for the three thresholds are shown in Table 4: 14.69%, 6.52%, and −1.54%, respectively. These three threshold values divide the sample set into four groups based on whether the value of the return on asset ratio (ROA) variable is less than or greater than the threshold estimates. Thus, the sample data are divided into four groups with ROA in the range ROA < −1.54, −1.54 ≤ ROA < 6.52, 6.52 ≤ ROA < 14.69, and 14.69 ≤ ROA.

4.2.2. Empirical Results

The model automatically segments the data into four groups for regression analysis based on detecting three thresholds. The outcomes of this process are presented in Table 5. Board members with accounting expertise exhibit a positive influence, particularly in the group where −1.54 ≤ ROA < 14.69, prominently in the 6.52 ≤ ROA < 14.69 range.

Table 5. Factors Affecting ROA at Different Groups.

Variables	ROA < −1.54 (Group 1)		−1.54 ≤ ROA < 6.52 (Group 2)		6.52 ≤ ROA < 14.69 (Group 3)		14.69 ≤ ROA (Group 4)	
Group	**[1]**		**[2]**		**[3]**		**[4]**	
C	−35.46	***	−1.85	**	10.93	***	16.22	***
ACC	−1.92		0.42		1.57	**	−1.79	
FIN	11.98	***	−1.72	**	−3.02	**	5.68	**
LAW	−10.58	***	−0.10		−2.11	*	3.36	
BDP	−0.01		0.01	***	0.00		0.04	***
MNP1	−0.13		−0.02		−0.02		0.14	***
MNP2	0.09	***	0.04	***	0.01		0.00	
BSZ	2.73	**	0.17		0.20		2.66	**
IBD	−2.66	***	0.19		−0.14		−0.38	
SIZE	4.28	***	0.70	***	−0.27	**	−0.61	**
DBT	−0.01		−0.01	*	−0.01	**	−0.03	***
NPT	0.01	***	0.02	***	0.05	***	0.18	***
Number of Obs.	384		2459		1930		536	
Group	4		3		2		1	
R^2:	0.839							
Adj.R^2:	0.837							
F-Statistic	583.342 ***							

Note: ***, **, and * represent the 1%, 5%, and 10% significant levels, respectively. ROA is the after-tax return on assets.

Conversely, board members with financial expertise demonstrate a clear positive impact on ROA when ROA is either below −1.54 or exceeds 14.69. On the other hand, board members with legal expertise predominantly have a detrimental effect on ROA, with the most significant impact observed when ROA is below −1.54%.

Board size (BDS) generally contributes positively to ROA, particularly in Groups 2 and 3. The positive influence of MNP1 on ROA is exclusively notable in Group 4. Meanwhile, the positive impact of MNP2 is observed across all groups and is especially significant when the ROA is less than 6.52.

The correlation between MNP1 and a positive effect on ROA is exclusive to Group 4, whereas MNP2 exhibits a positive influence across all groups, with heightened significance when ROA is below 6.52. A larger board size (BSZ) corresponds to a higher ROA. Conversely, a higher proportion of independent board members (IBD) correlates with a lower ROA, indicating that independent board members may still need to exert their control in this regard. It is worth noting that a larger company size (SIZE) does not necessarily lead to higher ROA, however, this is true when the debt (DBT) level is low.

In Group 1, where ROA values are higher (positive), if the coefficient for "board members with financial expertise" is positive, it implies that an increase in the presence of such board members is associated with a higher ROA, which is typically expected. In Group 4, where ROA values are lower (negative), if the coefficient for "board members with financial expertise" is negative (when interpreted inversely), it implies that an increase in the presence of such board members is associated with an even lower (more negative) ROA, suggesting that, in Group 4, board members with financial expertise may have a detrimental effect on ROA.

Group 3 poses a challenge because it encompasses a wide range of ROA values, including negative and positive ones. The coefficients for independent variables in Group 3 should be interpreted cautiously due to this mixed range of ROA values. The interpretation may vary depending on this group's specific values of ROA. When interpreting these coefficients, it is essential to consider the context and the actual data points.

The multifaceted impact of board members with financial and legal expertise on ROA underscores the importance of considering industry dynamics, economic conditions, and company-specific factors when interpreting these associations within different ROA groups. The interplay of these factors contributes to the complexity of understanding how board expertise influences a company's financial performance.

The problem requires predicting ROA from the board and accounting data, so the input layer contains the relevant values collected, and the output layer contains the ROA value from the network. Then, proceed with building a network or choosing a network structure by choosing the number of layers and the number of hidden neurons in each layer for the network.

Typically, an Artificial Neural Network (ANN) design starts with a single hidden layer. The number of neurons in this hidden layer is gradually adjusted until the network's output error reaches an acceptable level and the desired output value is achieved. A second hidden layer is introduced if the number of neurons is too large (exceeding 50) and the error remains unacceptable. This iterative process continues until the desired error rate and output accuracy are attained. In this study, we constructed a neural network with two hidden layers, each containing ten and six neurons, respectively. Figure 3 illustrates the architecture of the ANN used for predicting the ROA.

With the network structure selected, we proceed to the training step of the network. In essence, training the network is the process of adjusting the weights. These link weight values default to random at the beginning of network construction, then, during network training, the network algorithms adjust the above values.

Performance graph observation through network construction is used to evaluate the network and consider overfitting. The mean square error (MSE) is displayed at the loop location (Epoch) for the best efficiency of network construction.

Figure 4 depicts the error of the ANN model, showing that the predictive accuracy of the ROA of the model, including ACC, FIN, and LAW, is 78.475%, 78.783%, and 78.837%, respectively, compared with the actual data.

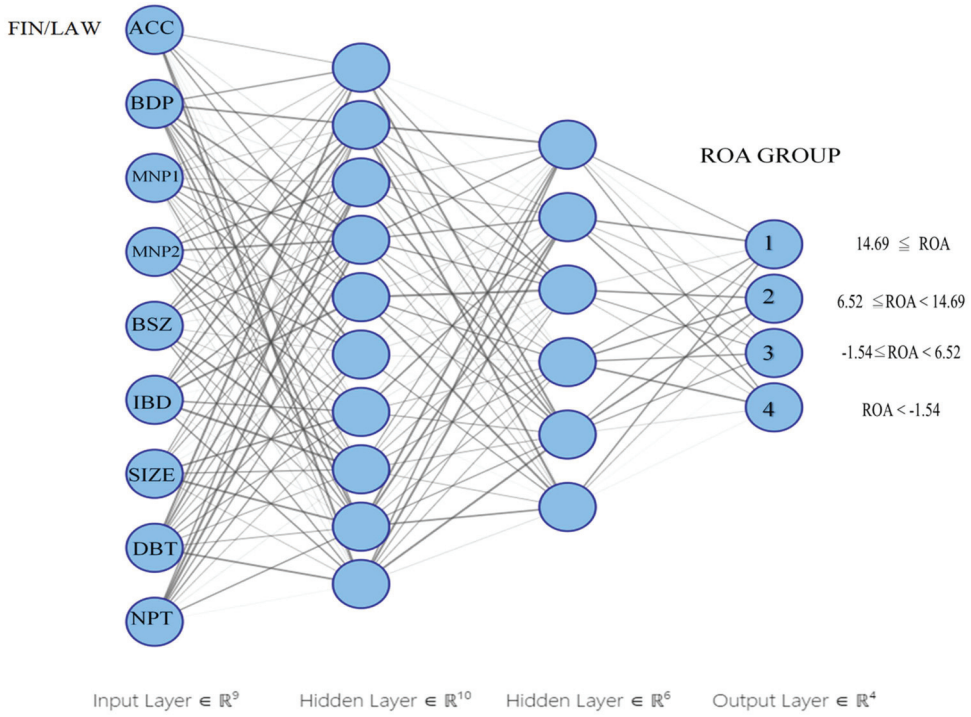
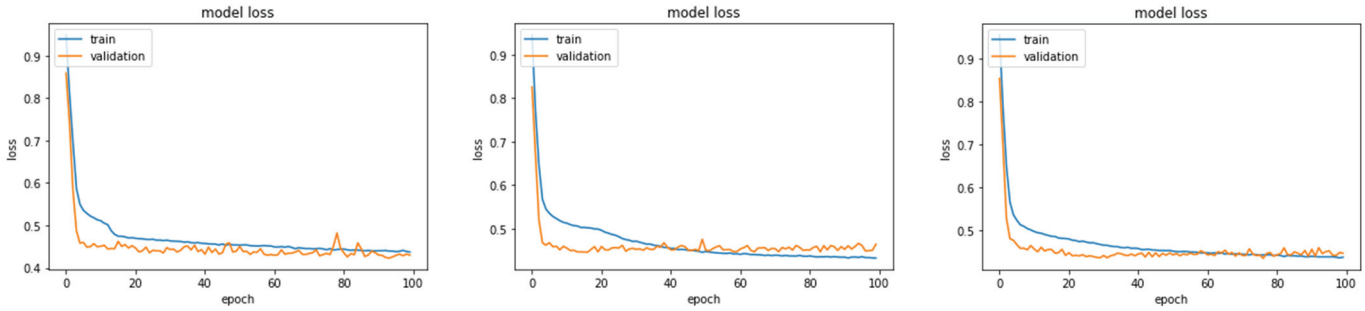


Figure 3. ANN for Predicting ROA.



The accuracy of ANN is 78.475
ACC

The accuracy of ANN is 78.783
FIN

The accuracy of ANN is 78.837
LAW

Figure 4. The model loss of the ANN models.

When the board of directors has more members with accounting qualifications, it will help to steadily increase the ROA value. In contrast, more members with financial expertise have a more extreme influence, which is not a stable contributor to ROA. Moreover, members with legal expertise do not have any significant positive influence. Directors holding more company shares can increase the ROA value only when the ROA is more excellent than 14.96%. However, the positive effect is more pronounced when considering the element of managers serving concurrently on the board of directors. Thus, the directors promote improving profits when managing and participating in the company's business decisions.

The larger the number of board members, the higher the ROA, however, the opposite is true with more independent board members. This result shows that independent members need to promote more of their obligations to control the activities of the companies. The

company's total assets only partially have a positive effect on ROA, especially for companies with a high ROA. At the same time, the debt ratio is purely the minus of ROA.

In the context of corporate governance and the impact of board characteristics on financial performance, this statement is not universally supported by theory alone. The relationship between the educational diversity of board members and financial performance is complex and multifaceted, and it can vary depending on various factors and contexts. Therefore, it is essential for empirical research, like the study mentioned in the question, to investigate and provide empirical evidence on the relationship between the educational diversity of board members and financial performance within a specific context or dataset. Empirical research can help validate or refute the theoretical notions and provide insights into the practical implications of board diversity for financial outcomes in real-world scenarios.

5. Conclusions

The research results show that, for companies in traditional industries that are listed on the Taiwanese stock market, the composition of the board of directors impacts the business's financial performance.

Boards with more members may have a greater capacity for overseeing corporate governance. A greater representation of board members with accounting expertise tends to lead to more consistent corporate profits, whereas an increased presence of members with financial expertise results in more pronounced profit fluctuations. Additionally, in the case of companies with lower profits, managers may face pressure from loan contracts or receive limited attention from investors, potentially leading to heightened practices related to profit management. Managers may aim to maximize their benefits. At this time, the return on assets is high, and the managers obtain the expected benefits. Thus, to restrain the profit management behavior of managers, enterprises need to have a broader view and have specific judgments and analyses in each enterprise to build organizational structure and procedures. The operation of the board of directors is more effective because of the industry's characteristics, the enterprise's size, the capital structure, and the performance of the enterprise. Finally, this study used deep learning and big data techniques to build an Artificial Neural Network (ANN) model that specializes in predicting the ROA of companies based on most factors related to board structure, with a prediction accuracy of over 78%.

While the research confirms a link between board composition and financial performance in traditional Taiwanese companies, the impact of specific expertise reveals intriguing nuances. Larger boards, per Hypothesis 1, might enhance decision-making capacity due to broader expertise, but communication challenges remain a concern. Hypothesis 2 is validated, as board members with accounting expertise promote stable profits through better financial oversight. Interestingly, Hypothesis 3's predicted substantial impact of financial expertise was found to likely lead to profit fluctuations, suggesting a focus on aggressive, short-term gains over long-term stability. This paper delves further, highlighting potential profit management practices in low-profit companies due to external pressure, emphasizing the need for closer monitoring. Notably, it introduces an Artificial Neural Network model capable of predicting a company's return on assets with impressive accuracy based on board structure factors. In conclusion, this research offers valuable insights by analyzing how different expertise and factors like board size impact performance, aiding companies in optimizing their board composition for informed decision-making and sustainable success.

Author Contributions: Conceptualization, C.-Y.L. and B.T.-T.-N.; methodology, B.T.-T.-N.; software, B.T.-T.-N.; validation, C.-Y.L., C.-R.W. and B.T.-T.-N.; formal analysis, C.-Y.L., C.-R.W. and B.T.-T.-N.; data curation, C.-R.W.; writing—original draft preparation, C.-Y.L., C.-R.W. and B.T.-T.-N.; writing—review and editing, C.-Y.L., C.-R.W. and B.T.-T.-N.; visualization, B.T.-T.-N.; supervision, C.-Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are sourced from the Taiwan Economic Journal (TEJ). After excluding non-available values, 5309 observations are used.

Conflicts of Interest: The authors declare no conflict of interest.

References

Adams, Renée D., and Daniel Ferreira. 2007. Board size and composition: A comparison of U.S. and U.K. listed firms. *Journal of Financial Economics* 85: 781–821.

Aguilera, Ruth V., Alvaro Cuervo-Cazurra, and Francesco Perrini. 2021. Board functional expertise and firm financial performance: The moderating role of economic downturns. *Strategic Management Journal* 42: 206–34.

Amini, Mohammad, and Fang Zhang. 2022. Industry-specific human capital and board oversight: Evidence from risk management effectiveness. *Journal of Accounting and Economics* 119: 101445.

Bai, Jushan, and Pierre Perron. 2003. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics* 18: 1–22. [CrossRef]

Boone, Linda E., and Maaike A. Lüdema. 2018. Boardroom diversity: Why do the board directors' gender and professional qualifications matter? *Review of Business and Economic Literature* 63: 349–89.

Carpenter, Mason A. 2002. The implications of strategy and social context for the relationship between top management team heterogeneity and firm performance. *Strategic Management Journal* 23: 275–84. [CrossRef]

Chang, Ruey-Dang, Chin-Jung Fang, and Yi-Cheng Tseng. 2007. Central agency problem and earnings management: An investigation of board structure and external monitoring mechanism. *Journal of Management* 24: 17–39.

Dass, Nishant, Omesh Kini, Vikram Nanda, Bunyamin Onal, and Jun Wang. 2014. Board expertise: Do directors from related industries help bridge the information gap? *The Review of Financial Studies* 27: 1533–92. [CrossRef]

DeFond, Mark L., K. Raghunandan, and K. R. Subramanyam. 2002. Do non–audit service fees impair auditor independence? Evidence from going concern audit opinions. *Journal of Accounting Research* 40: 1247–74. [CrossRef]

Fama, Eugene F., and Michael C. Jensen. 1983. Separation of ownership and control. *The Journal of Law and Economics* 26: 301–25. [CrossRef]

Francoeur, Claude, D. Lafond, and Marc Roy. 2021. Boardroom gender diversity and firm performance: A meta-analysis. *Academy of Management Journal* 64: 431–61.

Hakim, Muhammad Saiful, and Chih Liang Liu. 2021. Generalist or Specialist? The Skill of CEO and Director That Really Matter to Firm Performance. *Modern Economy* 12: 1768–81. [CrossRef]

Hambrick, D. C. 2005. Just How Much Does the CEO Matter? *Harvard Business Review* 83: 92–99.

Hambrick, Donald C., and Phyllis A. Mason. 1984. Upper echelons: The organization as a reflection of its top managers. *Academy of Management Review* 9: 193–206. [CrossRef]

Hansen, Bruce E. 1999. Threshold effects in non-dynamic panels: Estimation, testing, and inference. *Journal of Econometrics* 93: 345–68. [CrossRef]

Hansen, Bruce E. 2000. Sample splitting and threshold estimation. *Econometrica* 68: 575–603. [CrossRef]

Hlel, Khawla, Ines Kahloul, and Houssam Bouzgarrou. 2020. IFRS adoption, corporate governance and management earnings forecasts. *Journal of Financial Reporting and Accounting* 18: 325–42. [CrossRef]

Jensen, Michael, and Edward J. Zajac. 2004. Corporate elites and corporate strategy: How demographic preferences and structural position shape the scope of the firm. *Strategic Management Journal* 25: 507–24. [CrossRef]

Kim, Daehyun, and Laura T. Starks. 2016. Gender diversity on corporate boards: Do women contribute unique skills? *American Economic Review* 106: 267–71. [CrossRef]

Klein, April. 2002. Audit committee, board of director characteristics, and earnings management. *Journal of Accounting and Economics* 33: 375–400. [CrossRef]

Klein, Peter, and Philipp Zurbrügg. 2023. The contingent effects of board expertise and diversity on firm performance: A meta-analysis and review. *Journal of Management Studies* 60: 34–82.

Kor, Yasemin Y., and Chamu Sundaramurthy. 2009. Experience-based human capital and social capital of outside directors. *Journal of Management* 35: 981–1006. [CrossRef]

Krishnan, Ravi, and Fernando Martin Santos. 2008. Board composition and the acquisition of complementary knowledge: A theoretical and empirical exploration. *Strategic Management Journal* 29: 867–88.

Li, Mingxiang, and Pankaj C. Patel. 2019. Jack of all, master of all? CEO generalist experience and firm performance. *The Leadership Quarterly* 30: 320–34. [CrossRef]

Luo, Xin, Ming Yu, and Min Yan. 2020. Board age diversity and firm innovation: The moderating role of industry technological intensity. *Journal of Business Research* 116: 401–12.

Nguyen, T. Thanh Binh, and Qi-Wen Huang. 2020. Impact of gender and education on corporate social responsibility: Evidence from Taiwan. *Problems and Perspectives in Management* 18: 334. [CrossRef]

Pucheta-Martínez, María Consuelo, and Isabel Gallego-Álvarez. 2020. Do board characteristics drive firm performance? An international perspective. *Review of Managerial Science* 14: 1251–97. [CrossRef]
Rynes, Steven L., John A. Colbert, and Kenneth A. Brown. 2007. The role of human capital in corporate governance: Theoretical foundations and research directions. *Journal of Management* 33: 397–424.
Sun, Liang. 2021. Does the location of directors' additional positions matter? A new dimension of board structure. *Global Finance Journal* 49: 100648. [CrossRef]
Warfield, Terry D., John J. Wild, and Kenneth L. Wild. 1995. Managerial ownership, accounting choices, and informativeness of earnings. *Journal of Accounting and Economics* 20: 61–91. [CrossRef]
Weisbach, Michael S. 1988. Outside directors and CEO turnover. *Journal of Financial Economics* 20: 431–60. [CrossRef]
Wu, TsingZai C., Jeng-Ren Chiou, and Li Cheng. 2011. The Influence of Corporate Governance on a CPA's Going-concern Opinion for Financially Distressed Firms: An Empirical Study for Firms Listed in the Taiwan Stock Exchange. *Tai Da Guan Li Lun Cong* 21: 187.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

International Journal of
Financial Studies

MDPI

Article

Predicting Healthcare Mutual Fund Performance Using Deep Learning and Linear Regression

Anuwat Boonprasope [1,2] and Korrakot Yaibuathet Tippayawong [2,3,*]

[1] Program in Industrial Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai 50200, Thailand; anuwat_boonprasope@cmu.ac.th
[2] Supply Chain and Engineering Management Research Unit, Chiang Mai University, Chiang Mai 50200, Thailand
[3] Department of Industrial Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai 50200, Thailand
* Correspondence: korrakot@eng.cmu.ac.th; Tel.: +66-816-719-019

Abstract: Following the COVID-19 pandemic, the healthcare sector has emerged as a resilient and profitable domain amidst market fluctuations. Consequently, investing in healthcare securities, particularly through mutual funds, has gained traction. Existing research on predicting future prices of healthcare securities has been predominantly reliant on historical trading data, limiting predictive accuracy and scope. This study aims to overcome these constraints by integrating a diverse set of twelve external factors spanning economic, industrial, and company-specific domains to enhance predictive models. Employing Long Short-Term Memory (LSTM) and Multiple Linear Regression (MLR) techniques, the study evaluates the effectiveness of this multifaceted approach. Results indicate that incorporating various influencing factors beyond historical data significantly improves price prediction accuracy. Moreover, the utilization of LSTM alongside this comprehensive dataset yields comparable predictive outcomes to those obtained solely from historical data. Thus, this study highlights the potential of leveraging diverse external factors for more robust forecasting of mutual fund prices within the healthcare sector.

Keywords: financial modeling; mutual fund performance; multiple linear regression; deep learning; LSTM

1. Introduction

The capital market is a crucial source for mobilizing savings and providing long-term credit in Thailand (Wanaset 2018). It plays a significant role in the country's economic development. The capital market provides a platform for buying and selling securities such as stocks and bonds, where mutual funds actively participate by pooling funds from investors to invest in these securities. A mutual fund serves as an investment vehicle that aggregates funds from numerous investors, directing them towards a diversified portfolio encompassing various asset classes such as stocks, bonds, and other securities. The inherent advantages of mutual fund investment include risk mitigation through diversification, expert management by seasoned fund managers, and the convenience of daily liquidity, enabling investors to buy or sell shares on a daily basis. Opting for mutual funds proves to be an appealing choice for individuals looking to capitalize on diversified returns while benefiting from the expertise of professional fund management.

In the midst of the market volatility precipitated by the onset of the global COVID-19 pandemic in early 2020, central banks and ministries of finance across diverse nations initiated a series of policy interventions known as Quantitative Easing (QE). These interventions encompassed measures such as interest rate reductions, infusion of liquidity into the financial system, and the implementation of stimulus packages spanning various sectors throughout the period spanning 2020 to 2021. The primary objective of these policies was to

mitigate the adverse economic repercussions stemming from the COVID-19 crisis. However, the implementation of these measures inadvertently triggered an uptick in inflation rates across several nations in 2022, thereby prompting a swift and robust response from central banking authorities, including the U.S. Federal Reserve (Fed) and the European Central Bank (ECB). This response entailed a decisive tightening of monetary policy, characterized by aggressive interest rate hikes aimed at curbing inflationary pressures. Consequently, global equity markets experienced notable corrections, with returns on diverse asset classes registering downward adjustments commensurate with the prevailing market dynamics.

While 2022 witnessed pronounced market volatility, it is essential to note that not all sectors encountered uniform challenges. Notably, the healthcare sector demonstrated resilience and commendable performance amidst market fluctuations, attributable to several key factors. Firstly, healthcare stocks exhibited robust financial performance in recent periods, showcasing their ability to weather economic downturns and navigate high market volatility (Dillender et al. 2021). Secondly, external dynamics such as the COVID-19 pandemic, the aging global population, and advancements in medical technology significantly contributed to the sector's substantial growth. Thirdly, the healthcare sector maintains an appealing valuation, characterized by comparatively lower profit estimates in relation to other sectors, thus eliciting interest in exploring investment opportunities within healthcare securities. Given the favorable prospects for the healthcare industry, concerted efforts are underway to leverage and capitalize on the profit potential inherent in healthcare securities. One notable strategy involves the application of Machine Learning Models for forecasting future returns, thereby offering valuable insights to guide investment decisions within the healthcare sector.

Machine Learning (ML) has been increasingly integrated into investment strategies, enabling computer systems to process, predict, and make decisions independently through learning from input datasets (Alzubi et al. 2018; Janiesch et al. 2021). This approach empowers computers to autonomously handle and solve various problems by learning from the data fed into them. ML operates on principles similar to human learning, requiring the assimilation of experiences. In ML, the process involves feeding data and instructions to the computer for learning. To enhance outcomes, continuous input of new data is necessary, fostering consistent learning and model refinement. ML is categorized into three types: Supervised Learning, Unsupervised Learning, and Reinforcement Learning. Among these, Multiple Linear Regression (MLR) is a fundamental algorithm in supervised learning (Slinker and Glantz 1988). MLR requires data input for the model to learn, employing statistical calculations to produce numerical results. MLR finds applications in various fields, including the environment (Heshmaty and Kandel 1985), agriculture (Kittichotsatsawat et al. 2022), and finance (Alnabulsi et al. 2023), showcasing its versatility.

One of the most popular and widely discussed branches is Deep Learning (Kumar and Manash 2019). It is a mathematical model that mimics the workings of the neural networks in the human brain by combining multiple layers of neural networks into a highly complex architecture (Sarker 2021). This complexity makes it an efficient and highly accurate mathematical model. Deep Learning involves learning from sample data, and the acquired knowledge is then used for tasks such as pattern recognition, data categorization, or data prediction (Kumar and Manash 2019). Later developments in Deep Learning led to the creation of a mathematical model known as Long Short-Term Memory (LSTM).

LSTM, a recurrent neural network model tailored for time-series data analysis (Hochreiter and Schmidhuber 1997), demonstrates proficiency in handling vast datasets and decision-making, surpassing conventional artificial neural networks (Hochreiter and Schmidhuber 1997; Van Houdt et al. 2020). Modeled after the memory patterns of the human brain, LSTM possesses a constrained memory capacity, akin to the brain's process of discerning the significance of new events for acceptance or rejection (Hochreiter and Schmidhuber 1997). This distinctive architecture empowers LSTM to excel in capturing patterns from prolonged sequences (Bolboacă and Haller 2023; Hochreiter and Schmidhuber 1997; Van Houdt et al.

2020), rendering it well suited for analyzing time-series data, including historical stock prices (Ouyang et al. 2020; Gülmez 2023).

As mentioned earlier, investing in mutual funds involves risks stemming from the price volatility influenced by various factors (Banegas et al. 2022; Li 2020; Qureshi et al. 2017). In addition to the managerial capabilities in selecting investments in different fund units, external factors may also impact the fund's price volatility, particularly concerning economic issues (Kang et al. 2022). This study has therefore categorized the influencing factors into two groups: internal factors, involving the investment choices in various assets directly affecting the fund's price, and external factors, encompassing economic indicators reflecting market conditions and the country's economic state over time (Panigrahi et al. 2019). Both groups of factors are considered crucial and are diligently incorporated into the dataset to create a model capable of accurate and efficient predictions.

This study proposes the use of the MLR and LSTM model to forecast the trends in the prices of mutual funds in the healthcare sector in Thailand during the post-COVID-19 period. The approach involves utilizing external factors, which are economic indicators expected to influence the securities prices in the medical business sector based on previous studies. Additionally, internal factors such as past asset prices selected by the fund for investment are incorporated. Our study presents a paradigm shift in stock market prediction, going beyond the confines of historical trading data. The results illuminate the efficacy of incorporating various factors that influence the healthcare sector for accurate future price predictions. Notably, our exploration reveals that the application of Long Short-Term Memory (LSTM) models to this diverse set of data produces results on par with traditional methods reliant solely on historical data for forecasting. This breakthrough underscores the potential for a more robust and comprehensive approach to forecasting stock prices.

The above passage outlines the structure of the research study. It begins by highlighting the origin and significance of the identified gaps in previous literature and the introduction of machine learning tools. Following this, the study will proceed with a review of previous research, identification of gaps in the existing literature, and a comparative analysis of research outcomes similar to the current study. Subsequently, the methodology of the study will be comprehensively presented. The subsequent section will focus on a detailed discussion of the findings, including an exploration of the study's limitations. Finally, the study will be concluded by summarizing the results and suggesting potential directions for future research.

2. Literature Review

Brogaard and Zareei (2023) utilize machine learning algorithms to explore the profitability of technical trading rules based on historical stock prices. Their study confirms investors' ability to discover profitable rules through machine learning methods. Comparisons with other algorithms highlight evolutionary genetic algorithms' advantage in incorporating erroneous predictions, resulting in enhanced profitability. Evaluation across various periods consistently shows the selection of trading rules that perform well out of sample, maximizing abnormal returns. Additional tests on diverse datasets ensure the robustness of the findings. This research demonstrates the potential of utilizing machine learning in finance, particularly in employing complex and efficient models for computational tasks. The findings suggest that the methodologies and insights derived from this study can be extrapolated to other models, particularly those with high computational complexity and efficiency, for application in finance-related endeavors.

The utilization of machine learning, particularly in the form of deep learning models, has witnessed a notable surge within the realm of finance. Gu et al. (2020) elucidate that machine learning methodologies substantially augment empirical asset pricing frameworks, surpassing conventional regression-based methodologies. Their study delineates decision trees and neural networks as preeminent performers, adept at capturing intricate nonlinear interactions among predictors. A consensus emerges regarding the prominence of predictive signals such as momentum, liquidity, and volatility. These methodologies

proffer discerning insights for investors, potentially amplifying the efficacy of conventional strategies twofold, with a pronounced proficiency in forecasting returns for sizable, more liquid equities and portfolios. This underscores the burgeoning influence of machine learning within the fintech domain.

In a parallel vein, Zhou et al. (2023) employ deep neural network (DNN) models to predict the US equity premium, comparing their efficacy against ordinary least squares (OLS) and historical average (HA) models. The investigation reveals that DNN models consistently outshine OLS and HA counterparts across in-sample and out-of-sample assessments, alongside asset allocation simulations. Moreover, the integration of 14 supplementary variables sourced from finance literature bolsters the predictive accuracy of DNN models. Notably, the paper introduces a nonlinear machine learning paradigm for forecasting equity premiums, marking a departure from conventional econometric frameworks. Additionally, the study delineates the foundational equations underpinning the employed DNN models.

The points highlighted in the study review by Sonkavde et al. (2023) align with the recognition of deep learning models' prominence in the financial sector, particularly in stock price prediction and classification. The review underscores that deep learning models, with their capability to capture intricate patterns, handle extensive datasets, and engage in feature learning and representation, have gained popularity in forecasting and trend prediction for stock prices. Similarly, the findings from Shah et al. (2022), who discussed the limitations and accuracy of various models, including deep learning, support the notion that deep learning algorithms, such as LSTM, Convolutional Neural Networks (CNN), and their hybrid models, significantly impact stock prediction and portfolio management.

During the recent COVID-19 situation, there have been research efforts employing deep learning models to study forecasting trends. Ersin and Bildirici (2023) proposed the GARCH-MIDAS-LSTM model, which integrates LSTM deep neural networks with the GARCH-MIDAS model to predict stock market volatility. This research utilized data from the Borsa Istanbul stock market, specifically during the COVID-19 shutdown and economic reopening period in Turkey. An important aspect of this research is the incorporation of monthly explanatory variables, encompassing economic leading indicators such as the Composite Leading Index (CLI), the country-specific Geopolitical Risk Index (GPR) for Türkiye, and the cycle and trend industrial production indices (IPIC and IPIT). The findings indicate that stock market volatility is most effectively modeled with geopolitical risk, followed by industrial production, while the impact of future economic expectations is relatively lower. This demonstrates the capability of utilizing deep learning models during the COVID-19 situation and additionally highlights the integration of economic indicators in model development.

Similarly to Chimmula and Zhang (2020), who developed a Deep Learning forecasting model for COVID-19 in Canada utilizing LSTM networks for real-time predictions, this study demonstrates superior performance compared to other models. The model provides valuable insights into transmission rates across countries and serves as an alert system for frontline staff, aiding in crisis preparations. Key findings of the study include identifying a linear transmission trend in Canada, predicting an expected end within three months, and highlighting the model's effectiveness in guiding health authorities. The research underscores the impact of early social distancing measures and emphasizes the potential role of technology and international collaboration. In summary, the developed model presents a valuable tool for crisis management and prevention. However, all previous research works have not conducted in-depth studies in the healthcare sector.

In the domain of predicting healthcare stock prices, research employing machine learning techniques has indeed been conducted. Chatterjee et al. (2021) developed six models that integrated time series, econometric, and learning-based techniques. These models included Holt–Winters Exponential Smoothing, ARIMA, Random Forest, MARS, RNN, and LSTM. The objective was to forecast stock prices within three major sectors: IT, banking, and the healthcare sector. The research identified LSTM as the best-performing deep learning model, achieving a Root-Mean-Squared Error (RMSE) of 0.022 for the health-

care sector. Its proficiency in handling intricate sequential data, without encountering issues such as vanishing gradients and exploding gradients, contributed to the generation of highly accurate forecasts.

Similarly, in alignment with the work of Sen et al. (2021), which presents optimized portfolios based on the seven sectors of the Indian economy, including the health sector, the research utilized data spanning from 1 January 2016 to 31 December 2020. This research employed an LSTM regression model to forecast future stock prices and design optimized portfolios across the seven sectors. The paper specifically constructs an LSTM regression model for predicting future stock prices, and the projected returns and risks of each portfolio are computed five months after portfolio construction. The findings reveal the high accuracy of the LSTM model. However, it is noteworthy that both studies are comprehensive in their approach, forecasting prices across various industry sectors without a specific focus on the development of predictions within the healthcare sector.

Mokhlis et al. (2021) conducted a study that delved deeper into the forecasting development within the healthcare sector. In this research, the authors explored the historical trends of IHH healthcare stock by developing hybrid models, specifically ARIMA-GARCH and ARIMA-TGARCH. The investigation utilized data from September 2015 to September 2021, comparing the performance based on Root-Mean-Squared Error (RMSE) and Mean Absolute Error (MAE). The optimal hybrid model for forecasting IHH stock prices was identified as ARIMA (4,1,5)-GARCH (1,1), exhibiting superior accuracy with a smaller RMSE of 0.02289 and MAE of 0.01672. This research demonstrates results with RMSE values closely aligned with those obtained in the LSTM study conducted by Chatterjee et al. (2021).

The subsequent research by Jariyapan et al. (2022) focused on studying the nowcasting and forecasting of healthcare stock prices in the United States during the COVID-19 period, incorporating Google trend data. In the realm of machine learning, the research employed supervised learning algorithms, namely Linear Discriminant Analysis (LDA), k-Nearest Neighbors (kNN), and Support Vector Machine (SVM), to investigate the cycle regimes of healthcare stocks over the next five years. The research utilized data from five stock price indexes in the healthcare sector, coupled with Google trend data, spanning from 2015 to 2020. The findings from this research identified that LDA exhibited the highest coefficient validation. The results underscored that machine learning approaches, including clustering, classification, and parametric or nonparametric prediction, play a crucial role in econometrics. These approaches provide valuable information for investors to effectively manage their portfolios, particularly in the healthcare sector during the COVID-19 period.

Ahmed et al. (2022) conducted a comparison of the accuracies of various machine learning algorithms, including Linear Regression, Support Vector Regressor, Random Forest Regressor, and RNN with GRU. The authors selected the algorithm with the lowest Root-Mean-Squared Error (RMSE) value for the final model. The dataset used for this analysis comprised healthcare stock price data spanning the years 2016 to 2019. The research concludes that machine learning techniques, particularly RNN with GRU, which represents a single deep learning model among the considered algorithms, are effective for predicting healthcare sector stock prices. The chosen model achieved the lowest RMSE value of 0.051. This highlights the efficacy of deep learning methodologies in enhancing the accuracy of stock price predictions within the healthcare sector.

The collective research presented has explored the forecasting of healthcare sector securities using a variety of methods, including time series, econometric, and machine learning techniques. Nevertheless, the volume of studies is relatively limited, partly due to the recent rapid growth in the healthcare business in the preceding years. Previous research studies have identified gaps in academic literature, specifically: (1) There is a reliance on historical trading data to construct forecasting models. However, there has been a lack of research incorporating various factors influencing stock price volatility, despite extensive studies on such factors (Banegas et al. 2022; Li 2020; Qureshi et al. 2017). (2) Some studies (Sen et al. 2021; Mokhlis et al. 2021; Jariyapan et al. 2022) have focused on periods linked

to the COVID-19 situation, where market conditions were abnormal. However, the use of such data may not fully reflect the model's forecasting efficiency. (3) No identified research has delved into forecasting within healthcare mutual funds.

This study undertook an analysis of fundamental factors that may impact the performance and volatility of healthcare sector securities in three contexts, namely: (1) Economic Context: This involved an examination of the effects of economic policy on the healthcare sector, considering various economic indicators (Kang et al. 2022). (2) Industry Context: The study analyzed the state of the healthcare industry at both the national and global levels. It considered changes in prices for medical treatment and services within the country, as well as the dynamics of the global healthcare industry. (3) Company Context: The study delved into the performance of companies and the capabilities of their executives, reflected through the assets in which the funds invest. All three contexts are of concern and have been extended to influence the study of mutual fund price forecasting. Contexts (1) and (2) are considered external factors that impact the healthcare industry, while context (3) encompasses internal factors originating from the companies themselves, affecting the fund's performance.

Therefore, this study has introduced the use of Multiple Linear Regression (MLR) and Long Short-Term Memory (LSTM) methods to forecast the trend of mutual fund prices in the Thai healthcare sector during the post-COVID-19 period. The study utilized both internal and external factors, as mentioned earlier, for constructing forecasting models without relying solely on past price data. The objective is to present a model development that uses diverse data sources to demonstrate that various factors affecting the healthcare sector can be analyzed and utilized as inputs for predicting future prices. The results illustrate that incorporating a more diverse set of data beyond historical trading prices can enhance the effectiveness of forecasting models.

3. Materials and Methods

3.1. Data Collection and Descriptive Statistics

In this study, the price trading data of the Bualuang Global Health Care (BCARE) fund were selected for analysis. This fund invests solely in the feeder fund Wellington Global Health Care Equity Fund USD D Ac, which focuses on four subsectors: Major Pharmaceuticals, Biotechnology and Specialty Pharmaceuticals, Medical Products, and Health Services. The trading data of BCARE include daily closing prices, timestamped at the end of each trading day. The dataset spans from 21 December 2021, which corresponds to the date Thailand completed the administration of 100 million COVID-19 vaccine doses, to 30 October 2023, totaling 402 data points.

The dataset is divided into three segments: Training Data, Validation Data, and Test Data. Training Data and Validation Data combined constitute 80% of the dataset, with the remaining 20% designated as Test Data, resulting in 321 data points for Training and Validation, and 81 data points for Test Data. These segments are further divided at a 90:10 ratio, yielding 288 data points for Training Data and 33 for Validation Data from the initial 321. Figure 1 illustrates the dataset, showing the BCARE fund's closing price in Thai Baht over the study period.

In the part on external factors, this study has selectively chosen factors expected to impact the prices of mutual funds in the healthcare sector. General external influences encompass the SET50 Index, representing the top 50 companies of Thailand by average daily market capitalization, monthly inflation rates (Panigrahi et al. 2019; Cheng and Dewi 2020), the Consumer Confidence Index on a monthly basis (Bolaman and EVRİM 2014), quarterly GDP growth rates (Gyamfi Gyimah et al. 2021), and the exchange rate between the Thai Baht and the US Dollar (Jasra et al. 2012; Wong 2022). Moreover, specific factors tailored to healthcare sector funds include the monthly Consumer Price Index (Subhani et al. 2010; Jasra et al. 2012) for Health Care and Personal Care Services, along with the Dow Jones U.S. Health Care Index (Lin 2018). This comprehensive selection incorporates a total of 7 external factors in the analytical framework.

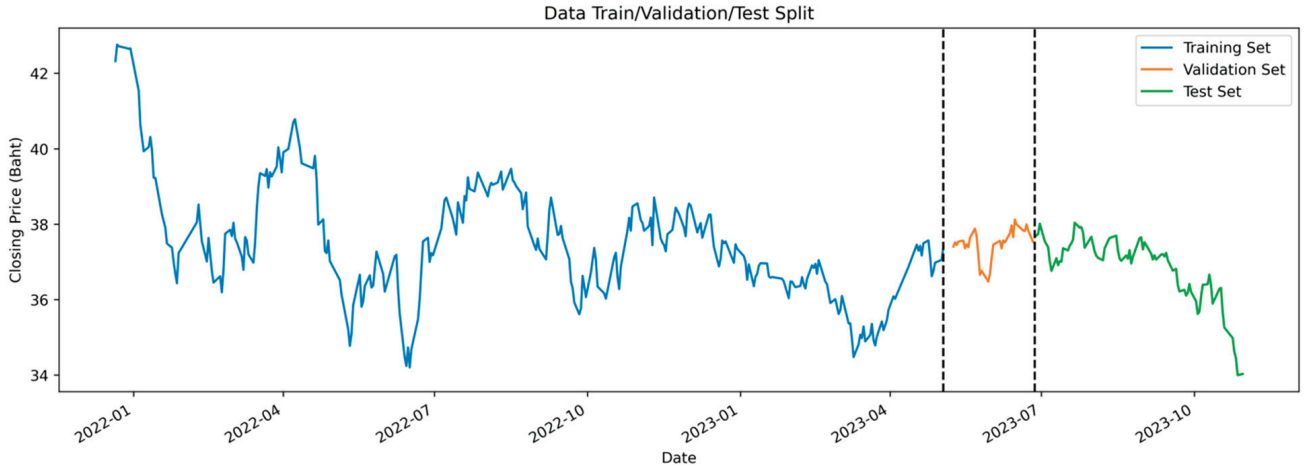


Figure 1. Daily closing price data of BCARE mutual fund.

In the section on internal factors, this study leverages historical daily price data of the top 5 holdings within the Wellington Global Health Care Equity Fund USD D Ac. These holdings consist of UnitedHealth Group Incorporated (UNH), Eli Lilly and Company (LLY), AstraZeneca PLC (AZN), Pfizer Inc. (PFE), and Danaher Corporation (DHR). A total of 12 factors (X_1, X_2, X_3, …, X_{11}, X_{12}) encompass both internal and external variables, with variable y representing the BCARE mutual fund price.

The training dataset encompasses the data utilized to facilitate the model's exposure and learning process. Subsequently, the validation dataset serves the purpose of evaluating metrics subsequent to the model's training phase, thereby assessing its performance and ascertaining optimal hyperparameters. Conversely, the test dataset is employed to appraise the model's efficacy in handling previously unseen data, thereby juxtaposing its predictions against actual values. In the context of this investigation, all 12 factors were employed for training purposes, with the model endeavoring to predict the variable 'y', denoting the price of the BCARE fund in Thai Baht. It is imperative to underscore that the prognosticated values within this study have undergone normalization. Detailed statistical insights and supplementary elucidations pertaining to each factor are delineated in Table 1.

Table 1. Statistical information and additional explanations for each factor.

Factors	Mean	Maximum	Minimum	SD	Description of Each Factor
BCARE (THB)	37.340	42.762	34.002	1.373	The historical data for the fund's prices, denoted as the variable 'y' for prediction by the model, are available on a daily basis and are stated in Thai Baht.
UNH (USD)	500.471	555.15	447.75	24.201	The historical stock price data for UnitedHealth Group Incorporated, which holds the top-ranking position within the fund's portfolio, are provided on a daily basis and are denominated in US dollars.
LLY (USD)	369.611	616.64	234.69	96.155	The historical stock price data for Eli Lilly and Company, the asset ranked second within the fund's portfolio, are provided on a daily basis and are denominated in US dollars.
AZN (USD)	10,593.664	12294	8282	905.072	The historical stock price data for AstraZeneca PLC, the asset ranked third within the fund's portfolio, are available on a daily basis and are denominated in US dollars.

Table 1. *Cont.*

Factors	Mean	Maximum	Minimum	SD	Description of Each Factor
PFE (USD)	44.538	59.55	30.11	6.968	The historical stock price data for Pfizer Inc., which is the fourth-ranked asset held within the fund, are provided on a daily basis and are denominated in US dollars.
DHR (USD)	247.868	328.47	185.1	30.178	The historical stock price data for Danaher Corporation, the asset ranked fifth within the fund's holdings, are available on a daily basis and are denominated in US dollars.
SET50 Index	965.682	1035.94	846.89	35.164	Index data referencing the top 50 highest-valued Thai stocks in the securities market, computed as a daily index.
US Dollars Exchange Rate (THB)	34.887	38.24	32.1	1.377	Daily exchange rate records detailing the conversion rate from US dollars to Thai Baht.
Dow Jones U.S. Health Care Index	1398.889	1540.53	1271.73	45.249	A market capitalization-weighted index that tracks the performance of the healthcare sector in the United States, presented on a daily basis.
Consumer Confidence Index	49.879	56.6	43.8	4.391	An economic indicator gauging consumer confidence and overall economic sentiment, including financial conditions. These data are reported on a monthly frequency.
Consumer Price Index for Health Care and Personal Care Services	102.345	103.6	100.71	0.948	The retail price index, which measures alterations in the prices of goods and services in equivalent quantities over a specified period, relative to the prices of the same commodities in the base year. This index specifically focuses on changes in the prices of medical treatment and services within the country. Monthly data are provided.
Inflation Rate	3.915	7.86	−0.31	2.691	The consumer price index, which quantifies the percentage increase in the general price level of goods and services within an economy over a specific period, reflecting the erosion of purchasing power of a currency. Monthly data are available.
Gross Domestic Product (GDP)	2.359	4.5	1.4	0.984	Gross Domestic Product (GDP), denoting the total monetary value of all finished goods and services produced within a nation's borders during a particular timeframe. This dataset is presented on a quarterly basis.

Figure 2 depicts the model-building procedure, which commences with data preprocessing to organize them for analysis. Subsequent to preprocessing, the data undergo normalization to ensure consistent scaling. Dimensionality reduction through PCA is then implemented to reduce data size and eliminate noise. The data are subsequently partitioned into training, validation, and test sets, specifically tailored for the LSTM model, whereas for the MLR model, it is divided into training and test sets at an 80:20 ratio. The ensuing steps entail training the data, fine-tuning hyperparameters using the validation set, and ultimately assessing the model's performance against the test set. The portions of X Test and y Test, which are segregated, represent out-of-sample data since they were not utilized in the model training process. This indicates that the model has not been exposed to or learned from this dataset previously. These segments are exclusively reserved for evaluating the performance of the trained and developed model.

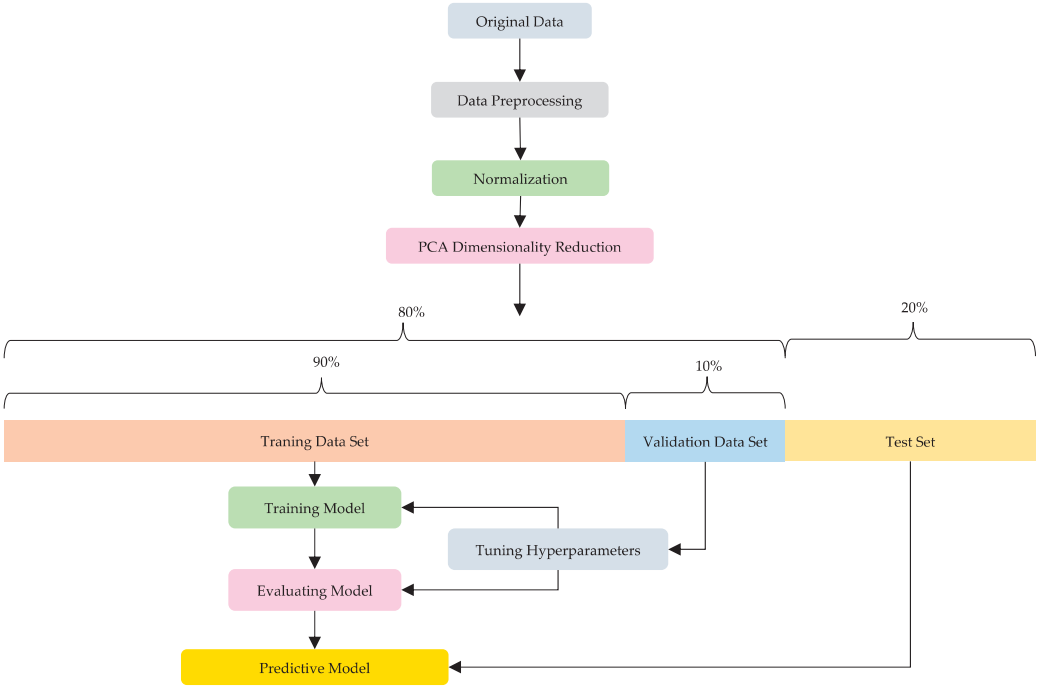


Figure 2. The workflow of the model-building process.

3.2. Principal Component Analysis

Principal Component Analysis (PCA) is a technique for reducing the dimensionality of data to facilitate analysis and conserve resources during model training. PCA achieves this reduction by projecting data vectors onto new axes called principal components. These components are chosen based on the variance observed along each axis. The PCA process involves three main steps (Jolliffe and Cadima 2016).

The first step is to compute the covariance matrix (C). The covariance matrix captures the relationships between the different features in the dataset, providing insights into how they vary together. This matrix is a critical input for the subsequent steps of PCA and is represented by Equations (1) and (2).

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{1}$$

$$C = \frac{1}{N}\sum_{i=1}^{N} (X_i - \mu)(X_i - \mu)^T \tag{2}$$

where N is the total number of data points, X_i represents the data matrix, where each row corresponds to a data point, and T denotes the transpose operation.

The second step involves finding the eigenvalue (λ) and eigenvector (V), both of which are components of the principal component and can be obtained from Equations (3) and (4).

$$Ax = b \tag{3}$$

where A is the transformation matrix or covariance matrix, x is the original vector, and b is the transformed vector.

$$Ax = \lambda x \tag{4}$$

where x is the eigenvector and λ is the eigenvalue.

The third step involves finding the weight vector (W) for each data point by projecting X_i onto the principal component axes ($V_1, V_2, \ldots, V_N$). The formula is presented in Equations (5) and (6).

$$W_k = V_k^T (X - \mu) \; ; \; k = 1, \ldots, N \tag{5}$$

$$W^T = [W_1, W_2, \ldots, W_N] \tag{6}$$

The results obtained from PCA analysis lead to the removal of less significant data, resulting in eigenvalues and eigenvectors. These two sets of data have corresponding relationships. When sorting eigenvalues in descending order, lower eigenvalues indicate less significant data.

3.3. Multiple Linear Regression

Multiple Linear Regression (MLR) involves data analysis to examine the relationship between a dependent variable (y_i) and multiple independent variables (X_i). It differs from Simple Linear Regression (SLR) in that MLR investigates relationships with more than one independent variable (Slinker and Glantz 1988). When there are k independent variables for a dependent variable, the MLR is presented in Equation (7).

$$y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_k X_{i,k} + \varepsilon_i \tag{7}$$

where y_i is the dependent variables, β_0 is the intercept, $X_{i,k}$ is the independent variables, β_k is the vector of slope, and ε_i is the random measured errors.

In the context of forecasting, especially within a time-series framework, the integration of a dynamic model incorporating lagged terms is imperative to capture temporal dependencies and enable accurate prediction. The MLR equation utilized for forecasting purposes is delineated in Equation (8).

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \ldots + \beta_k X_{k,t} + \gamma_1 Y_{t-1} + \gamma_2 Y_{t-2} + \ldots + \gamma_p Y_{t-p} + \varepsilon_t \tag{8}$$

where Y_t denotes the dependent variable at time t, while $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$ represent lagged values of the dependent variable up to p time periods prior. The coefficients $\gamma_1, \gamma_2, \ldots, \gamma_p$ correspond to the respective lagged terms.

3.4. Long Short-Term Memory

Long Short-Term Memory (LSTM) constitutes a variant of Recurrent Neural Network (RNN) architecture, conceived to offer heightened stability and efficacy (Hochreiter and Schmidhuber 1997). Notably, LSTM possesses the inherent capability to maintain the state or memory of individual nodes, thereby facilitating the retention of data origins and the retrieval of preceding values during backward temporal traversals. A distinguishing characteristic of LSTM lies in its incorporation of specialized gating mechanisms that regulate the flow of information into each node. These gating mechanisms include the Forget Gate Layer, Input Gate Layer, and Output Gate Layer, collectively facilitating nuanced information management within the network. The mathematical expressions governing the operations of each gate are explicated in Equations (9) and (11).

The forget gate layer

$$f_t = \sigma\Big(W_f.[h_{t-1}, x_t] + b_f\Big) \tag{9}$$

The input gate layer

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{10}$$

The output gate layer

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \tag{11}$$

where σ is sigmoid, W_x is the neuron gate (x) weight, h_{t-1} is the result of the preceding LSTM block, X_t is the input, and b_x is bias.

3.5. Data Preprocessing

In data processing, we often deal with different types of information that might have varying scales. Normalization and standardization are crucial steps in handling this diversity. They help ensure that all the data are on a similar scale, making them easier to compare and analyze. This is especially useful when dealing with variables that have widely different ranges, as these techniques ensure fair and consistent treatment across the board.

3.5.1. Normalization

Normalization is a method that adjusts data so that they fall within a scale of 0 to 1. It does this by subtracting the smallest value from each data point and then dividing it by the range of values (the difference between the maximum and minimum), as illustrated in Equation (12).

$$X_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{12}$$

where X_{scaled} is the normalized value, x is the original value, x_{min} is the minimum value of the features, and x_{max} is the maximum value of the features.

3.5.2. Standardization

Standardization is a process that makes data have a mean of 0 and a standard deviation of 1. It achieves this by subtracting the mean from each data point and then dividing the result by the standard deviation, as illustrated in Equation (13).

$$Z = \frac{x - \mu}{\sigma} \tag{13}$$

where Z is the standardized value, x is the original value, μ is the average value of the features, and σ is the standard deviation of the features.

3.6. Performance Metrics

This study utilized evaluation metrics, including Root-Mean-Squared Error (RMSE), Mean-Squared Error (MSE), and Mean Absolute Error (MAE), to compare the performance of the LSTM mutual fund prediction model and assess its effectiveness. All of the performance metrics are mathematically represented in Equations (14)–(16).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{14}$$

$$MSE = RMSE^2 \tag{15}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{16}$$

where y_i denotes actual value, $\hat{y}_i$ denotes predicted value, and $\overline{y}_i$ denotes the mean of y_i value.

3.7. Diebold–Mariano Test

The Diebold–Mariano test serves as a statistical method for comparing the forecast accuracy between two models, designated as Model 1 and Model 2 (Diebold and Mariano 1995). The test statistic, denoted as DM, is calculated as the difference in mean-squared forecast errors (DMSFE) divided by the standard error of the differences, as depicted in Equation (17).

$$DM = \frac{DMSFE}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{e_{1t} - e_{2t}}{T}\right)^2}} \tag{17}$$

where e_{1t} and e_{2t} represent the forecast errors of Model 1 and Model 2, respectively, at time t, while T signifies the total number of observations. This test aims to ascertain whether there exists a statistically significant distinction in forecast accuracy between the two models.

When the Diebold–Mariano (DM) statistic significantly deviates from zero, it signifies that one model demonstrates superior performance compared to the other. The corresponding p-value offers insight into the significance level of this discrepancy, thereby facilitating informed decision-making regarding model selection.

4. Results and Discussion

4.1. Dimensionality Reduction

Principal Component Analysis (PCA) is a method used to reduce the dimensionality of large datasets by transforming numerous features or X values into a smaller set that still retains the essential information of the dataset (Jolliffe and Cadima 2016). This is particularly useful for datasets with a large number of features, as it simplifies exploration and visualization, making data analysis more efficient. Additionally, working with a smaller dataset helps avoid issues like overfitting, where models may try to capture noise in the data, leading to improved model generalization.

In the initial step of PCA, the process begins with standardizing the entire dataset to ensure that each feature has an equal impact on data analysis. Subsequently, the covariance matrix is computed, representing the covariance values between all possible pairs of features in the dataset. A positive covariance indicates a direct relationship, implying that the variables increase or decrease together (correlated). Conversely, a negative covariance signifies an inverse relationship, suggesting that when one variable increases, the other decreases (inversely correlated). This covariance matrix provides insights into the relationships among different features in the dataset.

In the final step, the covariance matrix is used to calculate eigenvectors and arrange them in descending order based on their corresponding eigenvalues. This process allows us to identify principal components in order of importance. At this stage, a choice can be made to either retain all components or discard less significant ones (those with lower eigenvalues). Table 2 presents the eigenvalues for each component and cumulative values, illustrating how well the selected components cover the variance of the entire dataset. It is evident that choosing to retain the first 6 components covers approximately 96.23% of the dataset's variance, exceeding the 95% threshold. Thus, this study opted to reduce the dimensionality of features to only 6 dimensions. However, it is crucial to note that this dimensionality reduction does not involve discarding data but rather constructing new features that effectively summarize the existing information. Table 3 displays the covariance values for all 12 original features and their relationships with the newly created 6 components.

Table 2. The eigenvalues for each component and cumulative values.

Principal Component	Explained Variance	Explained Variance Ratio	Cumulative Explained Variance Ratio
1	6.27746	0.52182	0.52182
2	2.37639	0.19754	0.71936
3	1.46909	0.12211	0.84147
4	0.76288	0.06341	0.90489
5	0.44540	0.03702	0.94191
6	0.24519	0.02038	0.96230
7	0.17680	0.01469	0.97699
8	0.09612	0.00799	0.98498
9	0.07860	0.00653	0.99152
10	0.05038	0.00418	0.99570
11	0.03661	0.00304	0.99875
12	0.01501	0.00124	1.00000

Table 3. The covariance values for original features with the newly created 6 components.

Factors	PC1	PC2	PC3	PC4	PC5	PC6
UNH	−0.05395	0.39391	−0.58507	0.2138	0.11052	−0.0905
LLY	0.35898	0.01049	−0.17060	0.31135	0.02323	−0.33053
AZN	0.24575	0.11064	−0.36409	−0.67845	0.22114	0.18960
PFE	−0.37475	−0.02941	−0.13328	−0.06051	0.23523	0.27940
DHR	−0.36112	0.02441	−0.20098	0.22107	−0.13946	−0.11840
SET50	−0.32792	−0.13587	−0.11733	−0.32707	−0.38088	−0.65376
USD	0.11597	0.58036	−0.03926	0.25511	0.048396	−0.05818
DJUSHC	−0.10143	−0.35613	−0.60917	0.13878	−0.27575	0.31157
CCI Index	0.37847	−0.09041	−0.11985	−0.13022	−0.20164	0.01365
CPI Index	0.37481	0.06522	−0.13017	−0.16696	−0.16590	−0.24640
Inflation Rate	−0.32748	0.26784	−0.03579	−0.28068	0.36433	−0.2438
GDP Growth	−0.12131	0.51451	0.13735	−0.17183	−0.6627	0.32604

4.2. MLR Prediction Results

In the context of MLR, generally, it is necessary to satisfy the assumptions of multiple linear regression before performing MLR to ensure reliable results. One of these assumptions is the absence of multicollinearity, meaning that none of the predictor variables should be highly correlated with each other. Conventionally, an analysis of correlation values extracted from the correlation matrix, depicted in Figure 3, is conducted to assess the relationships between the dependent and independent variables. The correlation values range from -1 to 1, and a correlation exceeding 0.8 indicates a high level of correlation between variables (Berry and Feldman 1985). If such multicollinearity exists, it can impact the accuracy of various statistical estimates.

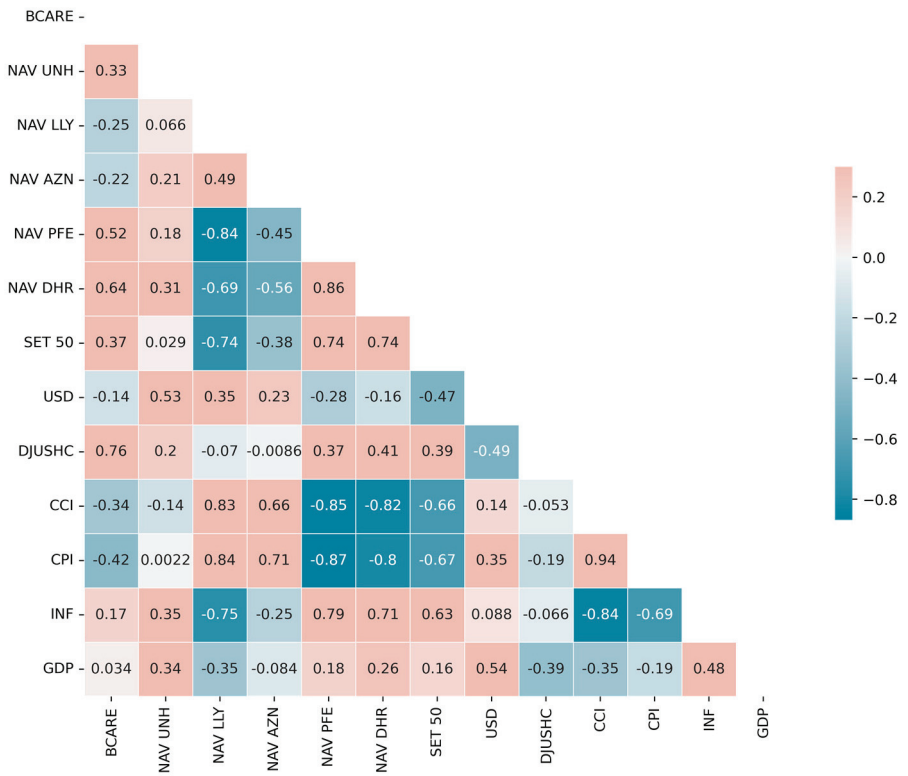


Figure 3. Heatmap correlation matrix of 12 feature.

However, this study has proposed the use of all 12 features in creating MLR. The data are divided into Training Data and Test Data in an 80:20 ratio, with no separate Validation Data. The outcomes of employing MLR for prediction are delineated in Table 4 and

Figure 4. Within the *y*-axis section, representing targets and output, the graph illustrates the model's predicted prices compared with the actual values, while the *x*-axis denotes the total number of data points. It is observed that the MLR demonstrates efficiency in predicting the Train dataset interval with an MSE of 0.3119 and RMSE of 0.5585. However, the model tends to exhibit characteristics of attempting to fit noise data excessively during this interval. Consequently, during the Test dataset interval, MLR predicts results with significantly reduced effectiveness, as evidenced by an MSE of 2.0046 and RMSE of 1.4158. The outcomes of utilizing MLR reveal a notable issue of overfitting, signifying a scenario where the model is trained to be overly complex and, consequently, cannot be effectively applied when encountering new data. The performance metrics for MLR prediction include an RMSE Overall of 0.8081 and MSE Overall of 0.6530. These findings emphasize the challenge of overfitting in the MLR model, illuminating its constraints when encountering unfamiliar datasets.

Table 4. RMSE and MSE for MLR prediction mutual fund prices.

Training Set		Testing Set	
RMSE Train	**MSE Train**	**RMSE Test**	**MSE Test**
0.5585	0.3119	1.4158	2.0046

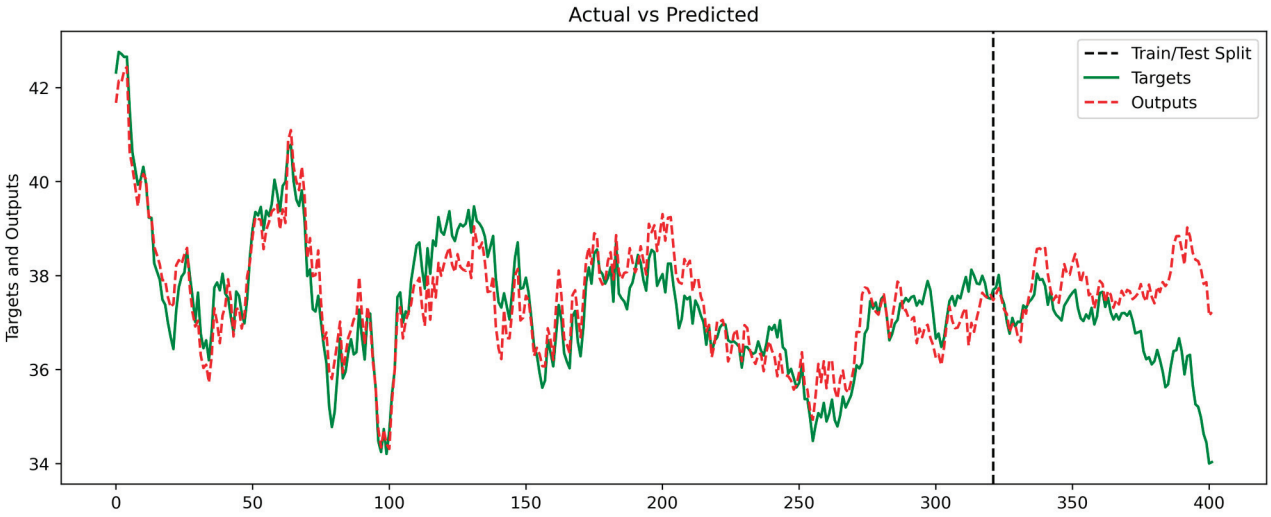


Figure 4. Targets and outputs for MLR prediction mutual fund prices.

4.3. LSTM Prediction Results

In the section on model tuning, this study introduced the tuning of the number of neurons and hidden layers. Figure 5 illustrates the model architecture designed for this purpose. The tuning focused on layers 1, 2, and 3, with variations in the number of neurons. Furthermore, LSTM layers and Dense layers were added after the initial tuning of the specified layers. The number of neurons considered for tuning included 32, 64, 128, and 256. The input data provided to the input layer encompass all 12 factors mentioned earlier. The optimized number of neurons is implemented across all layers, including the input layer, as illustrated in Figure 5, with the exception of the output layer, which consists of only one neuron. Moreover, we have established the duration of observations considered by the model when learning a time series, commonly referred to as the window size. This determination was influenced by the relatively restricted number of price data points available post-COVID-19. In this investigation, various window sizes were tested, namely 10, 12, 15, and 20 days. The batch size was set at 64, and epochs were configured to 40. To mitigate overfitting, Dropout layers were incorporated after the Input Layer and every Hidden Layer, each with a dropout rate of 0.2.

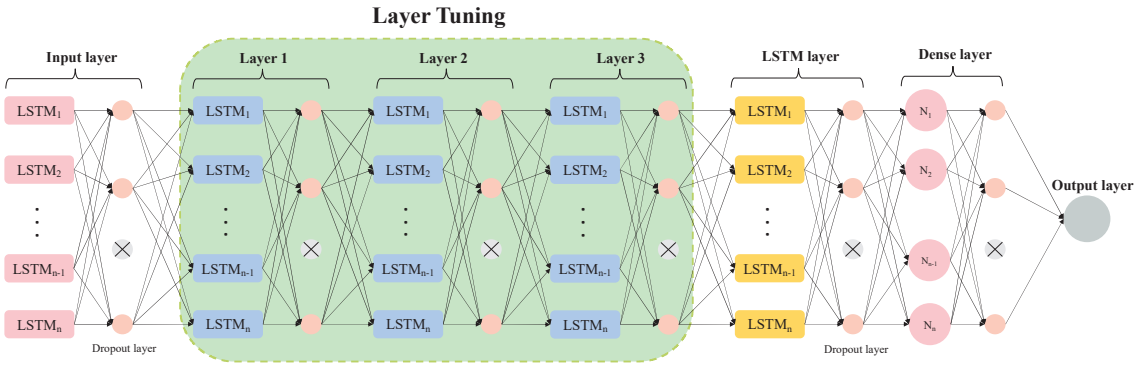


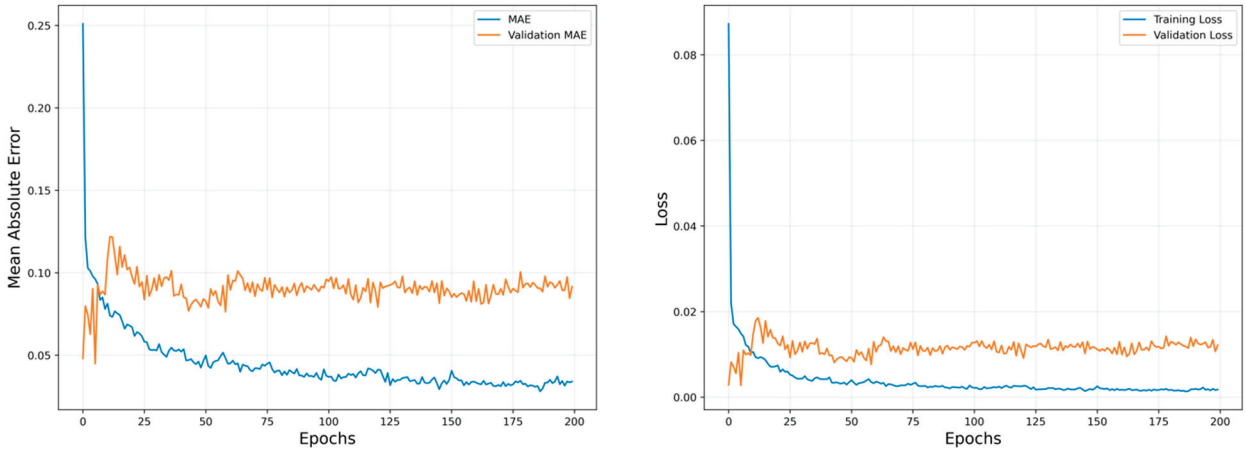
Figure 5. The architecture of the LSTM model in this study.

Table 5 presents the top 10 outcomes from a pool of 32 distinct configurations, showcasing the lowest Mean-Squared Error (MSE) values achieved through the tuning of neuron numbers and hidden layers across various window sizes. Notably, with a window size of 10 days, the recorded performance metrics reveal an MSE of 0.00301 for the training dataset and an MSE of 0.00942 for the validation dataset.

Table 5. Prediction results under different number of neurons, hidden layers, and window sizes.

Window Size	LSTM Layer 1	LSTM Layer 2	LSTM Layer 3	Number of Neurons	MSE Train	MSE Validation
10 days	1	1	1	256	0.00301	0.00942
	1	0	1	256	0.00361	0.01004
	1	1	0	256	0.00619	0.01089
	1	1	0	32	0.01047	0.01149
	1	1	1	64	0.00383	0.01175
	0	1	1	64	0.00379	0.01199
	0	1	1	256	0.00534	0.01210
	0	0	1	256	0.00372	0.01246
	0	1	0	2256	0.00310	0.01275
	0	1	1	32	0.00459	0.01426
12 days	1	1	1	128	0.00347	0.00954
	0	1	1	256	0.00370	0.01039
	1	1	0	256	0.00306	0.01288
	1	1	1	256	0.00302	0.01298
	1	0	1	256	0.00380	0.01328
	1	1	1	32	0.00502	0.01363
	1	1	1	64	0.00369	0.01440
	0	0	1	256	0.00613	0.01512
	0	1	0	64	0.00443	0.01523
	1	0	0	128	0.00537	0.01579
15 days	1	1	1	128	0.00409	0.01394
	1	1	1	256	0.00452	0.01467
	1	1	1	64	0.00353	0.01623
	1	1	0	256	0.00380	0.01684
	1	0	1	256	0.00470	0.01734
	0	0	1	256	0.00375	0.01774
	1	1	0	32	0.00428	0.01803
	0	0	0	256	0.00421	0.01954
	0	1	1	64	0.00315	0.01978
	1	1	1	32	0.00421	0.02021
20 days	1	1	1	256	0.00329	0.01175
	1	1	0	256	0.00397	0.01207
	0	1	1	256	0.00322	0.01450
	1	1	1	128	0.00293	0.01451
	1	0	1	256	0.00445	0.01461
	1	1	1	64	0.00338	0.01552
	0	1	0	256	0.00517	0.01670
	1	0	0	256	0.00516	0.01721
	0	0	1	256	0.00465	0.01816
	0	0	0	256	0.00416	0.01876

The Learning Curve is a graph illustrating the performance of the model on both Training Data and Validation Data, measured after the hyperparameter tuning process. It aims to identify whether the model suffers from issues like overfitting or underfitting (Anzanello and Fogliatto 2011). The x-axis represents the number of training cycles (Epochs), while the y-axis shows the model's performance. Figure 6a illustrates both the Mean Absolute Error (MAE) and Training Loss. After obtaining the optimal hyperparameters, they exhibit a continuous decreasing trend until approximately 25 epochs, where they stabilize. Similarly in Figure 6b, Validation MAE and Validation Loss increase initially and then stabilize as the model is trained for more epochs. The observed pattern in the Learning Curve indicates a well-fitting model, demonstrating good learning capabilities and the ability to generalize to unseen data.



(**a**) (**b**)

Figure 6. Learning curve performance: (**a**) Mean Absolute Error curve, (**b**) Loss curve.

Table 6 shows the results obtained from the predictions. The y-axis represents both targets and output, denoting the prices predicted by the model in comparison to the actual values. Meanwhile, the x-axis indicates the total number of data points. It is important to note that the values on the y-axis represent prices after normalization, scaled between 0 and 1. It can be observed that the model learns and predicts well within the range of the training data, with an RMSE of 0.0617 and MSE of 0.0038. The use of a 10-day window size for predicting future prices allows the model to forecast trends rather than capturing noise in the data. The model's accuracy on the validation set slightly decreased from the training phase, yielding an RMSE of 0.0458 and MSE of 0.0021. Conversely, during the testing phase, there was an increase in error metrics compared to before, with an RMSE of 0.0547 and MSE of 0.0030. The model accurately predicts a significant downward trend in future data, consistent with the actual test data showing a decline in fund prices as illustrated in Figure 7. Overall, this model produces an RMSE Overall of 0.0596 and MSE overall of 0.0035.

Table 6. RMSE and MSE for LSTM prediction mutual fund prices.

Training Set		Validation Set		Testing Set	
RMSE Train	**MSE Train**	**RMSE Validation**	**MSE Validation**	**RMSE Test**	**MSE Test**
0.0617	0.0038	0.0458	0.0021	0.0547	0.0030

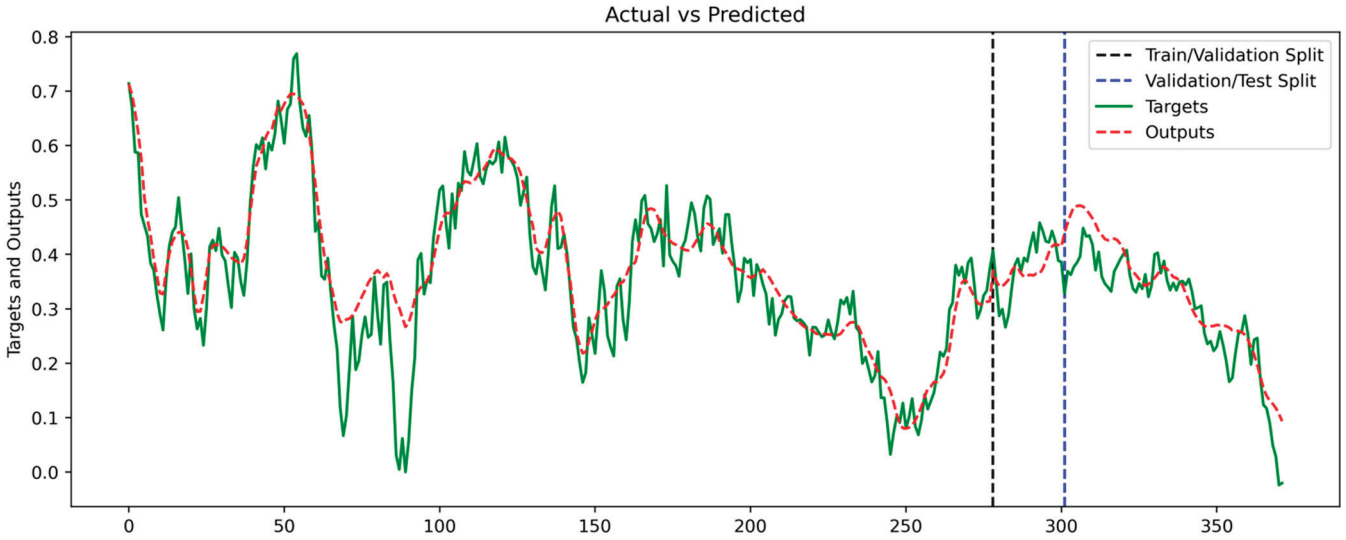


Figure 7. Targets and outputs for LSTM prediction mutual fund prices.

4.4. Diebold–Mariano Test

The Diebold–Mariano test was utilized to evaluate whether there is a statistically significant difference in forecast accuracy between LSTM and MLR, as detailed in Table 7. In this analysis, the Diebold–Mariano test statistic was computed as −2.2334. The negative value indicates the difference in mean-squared forecast errors between the two models, adjusted for the variance of the differences, suggesting that LSTM demonstrates a lower mean-squared forecast error compared to MLR. Regarding the *p*-value, it was found to be 0.02867. This value represents the probability of observing a test statistic as extreme as the calculated one under the assumption that the null hypothesis is true. With a *p*-value less than 0.05, the observed difference is deemed statistically significant in this study.

Table 7. The results of the Diebold–Mariano Test.

	Diebold–Mariano Test Statistic	***p*-Value**
DM test based on MLR and LSTM	−2.2334	0.02867

In the comparative analysis with previous studies on forecasting healthcare securities, as presented in Table 8, which are relatively limited in quantity, this research demonstrates a favorable RMSE value of 0.0547. This result surpasses the performance of Linear Regression, SVM, and Random Forest models, and in comparison to the LSTM model, the findings are closely aligned. However, it is noteworthy that the RMSE obtained in this study, at 0.0547, is slightly higher than the RMSE reported in the previous study by Ahmed et al. (2022), where the RMSE was 0.051. Nevertheless, it is important to acknowledge that the results of this research fall short compared to the studies conducted by Chatterjee et al. (2021) and Mokhlis et al. (2021).

In this study, our aim is to demonstrate the utilization of diverse influencing factors from economic, industrial, and corporate contexts to forecast future price trends in the healthcare industry. The results obtained indicate an enhancement over using general historical trading data in certain machine learning models (Ahmed et al. 2022). Although the outcomes may not surpass those of previous studies (Chatterjee et al. 2021; Mokhlis et al. 2021), we introduce a novel data approach that extends beyond solely relying on past securities trading data. By examining the future price prediction of healthcare mutual funds using various external factors, the results closely align with historical data, suggesting the potential application of these external factors in forecasting securities across other sectors. It is important to note that these factors may vary across different industries.

Transforming diverse data into analyzable formats could facilitate the integration of these factors into accurate predictive models, thereby contributing to improved forecasting across various industries.

Table 8. Comparison of healthcare securities prediction performances with the literature.

References	Subject	Description of Data	Model	RMSE	Accuracy
Ahmed et al. (2022)	The paper incorporates various machine learning algorithms, including SVM, reinforcement learning, ANN, and RNN, to forecast stock prices within the healthcare sector.	The dataset encompasses healthcare stock price data spanning the years 2016 to 2019, comprising fields such as opening and closing prices, alongside features such as price volatility and momentum.	Linear Regression	0.080	-
			RNN with GRU	0.051	-
			SVM	0.079	-
			Random Forest	0.065	-
Jariyapan et al. (2022)	Supervised learning algorithms such as Linear Discriminant Analysis (LDA), k-Nearest Neighbors (kNN), and Support Vector Machine (SVM) are employed to explore the cycle regimes of healthcare stocks over the next five years.	Monthly stock price data from 2015 to 2020 for five healthcare sector stock price indexes, specifically sourced from the Nasdaq index, were utilized in the paper.	LDA	-	0.8138
			k-NN	-	0.5223
			SVM	-	0.7847
Chatterjee et al. (2021)	Six models are developed, integrating time series, econometric, and learning-based techniques, specifically tailored for stock price prediction across three major sectors, with a particular focus on the healthcare sector.	Data pertaining to SUN Pharmaceuticals, covering the period from January 2004 to December 2019, were employed in the study.	Holt–Winters	0.056	-
			ARIMA	0.020	-
			Random Forest	0.009	-
			MARS	0.017	-
			RNN	0.0209	-
			LSTM	0.022	-
Mokhlis et al. (2021)	Time series models such as ARIMA, GARCH, and TGARCH are utilized to predict the IHH stock price, and their performances are evaluated using RMSE.	The paper leverages daily data of the IHH stock price to forecast its future trends and volatility, encompassing the period from September 2015 to September 2021.	ARIMA (4,1,5) -GARCH (1,1)	0.02289412	-
			ARIMA (4,1,5) -TGARCH (1,1)	0.02289852	-

Regarding the limitations of this study, it is possible that there are factors and data points overlooked beyond what has been presented, potentially extending beyond the scope analyzed in this study. These could encompass additional economic indicators, insightful data within the healthcare industry, or other contexts that can be quantified for analysis. It is advisable to consider incorporating such data to enhance the comprehensiveness of the analysis. Furthermore, the combination of models has the potential to improve accuracy and mitigate risks associated with individual model limitations, ultimately leading to more reliable predictions.

In addition to the aforementioned points, there is also the issue of applying this model in practical usage. Forecasting the value of y in the actual future necessitates knowledge of the values of X in the future. Specifically, this entails knowing the values of X_{t+1}, X_{t+2}, . . ., X_{t+n}. Thus, there is a need to consider making X dynamic through a process known as rolling forecast in machine learning. This can be achieved by constructing a model from the existing X data. Such models may include regression models, time series models, machine learning algorithms, deep learning models, or other models that are suitable for the dataset to predict the values of X in the future. Subsequently, the predicted values of X are used to forecast the value of y in an LSTM model that has undergone hyperparameter tuning. However, it is imperative that the models used for predicting each X value exhibit efficacy, as the accuracy of predicting X directly influences the prediction of y.

5. Conclusions and Future Work

In conclusion, this study has presented the utilization of both internal and external factors expected to impact the prices of mutual funds in the medical business sector. These factors were used to build a model for predicting future price trends. The external factors include the SET50 Index, inflation rates, Consumer Confidence Index, GDP growth rates, exchange rates between the Thai Baht and the US Dollar, Consumer Price Index for Health Care and Personal Care Services, and Dow Jones U.S. Health Care Index. Additionally, internal factors consist of historical daily price data for the top 5 holdings: UnitedHealth Group Incorporated (UNH), Eli Lilly and Company (LLY), AstraZeneca PLC (AZN), Pfizer Inc (PFE), and Danaher Corporation (DHR), resulting in a total of 12 features. This study has presented the selection of data to cover various factors that may affect mutual fund prices in the healthcare sector only. However, in other sectors, there may be other factors affecting fund prices beyond what has been discussed in this study. Additionally, the data used are relatively limited in quantity.

This study employed PCA to reduce the dimensionality of the data, making them more manageable for faster processing and avoiding overfitting issues associated with capturing data noise. The data dimensionality was reduced from 12 features to 6 features, retaining up to 96.23 percent of the information.

In the part on MLR, its predictive performance was less effective during the testing phase. The MLR model yielded results with an RMSE Test of 1.4158 and MSE Test of 2.0046, suggesting limitations in predicting outcomes during scenarios involving unseen data.

And in the LSTM section, hyperparameter tuning was conducted, resulting in the optimal configuration of a 4-layer LSTM followed by 1 dense layer, featuring 256 neurons, a batch size of 64, 40 epochs, a dropout rate of 0.2, and a specified window size of historical data for the past 10 days. The predictive results yielded an RMSE Test of 0.0547 and MSE Test of 0.0030. In addition, an analysis utilizing the Diebold–Mariano test has shown a statistically significant difference in the prediction results between the two, with LSTM exhibiting a lower MSE than MLR. The reduction percentages for RMSE and MSE when using LSTM are approximately 96.13% and 99.85%, respectively.

This study demonstrates that utilizing both internal and external factors in conjunction with LSTM is more effective in forecasting trends in healthcare sector mutual fund prices compared to MLR. The price fluctuations in these funds are influenced by various contributing factors, and these diverse elements can be used as valuable data for constructing predictive models to anticipate future trends.

For future work, we plan to explore the application of other machine learning models such as Decision Trees, Random Forests, or Artificial Neural Networks. Additionally, we aim to investigate model combinations such as LSTM-GRU, LSTM-CNN, and LSTM-VAR. Our future plans involve expanding the findings by incorporating additional factors beyond those presented in this study and integrating them with historical trading data. This comprehensive approach will be extended to mutual funds across various sectors. Furthermore, we also contemplate applying this model for rolling forecast by constructing various models to predict the value of X, including regression models, time series models, machine learning algorithms, deep learning models, or other models suitable for the dataset. These models are employed to forecast the value of y in the future using LSTM.

Author Contributions: Conceptualization, K.Y.T.; methodology, A.B.; software, A.B.; validation, K.Y.T.; formal analysis, A.B.; investigation, A.B.; resources, K.Y.T.; data curation, A.B.; writing—original draft preparation, A.B.; writing—review and editing, K.Y.T.; visualization, A.B.; supervision, K.Y.T.; project administration, K.Y.T.; funding acquisition, K.Y.T. All authors have read and agreed to the published version of the manuscript.

Funding: This study was partially supported by the project 'A Strategic Roadmap Toward the Next Level of Intelligent, Sustainable and Human-Centered SME: SME 5.0' from the European Union's Horizon 2021 research and innovation program under the Marie Skłodowska-Curie Grant agreement No. 101086487.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data were collected from https://www.investing.com (accessed on 1 November 2023).

Acknowledgments: This work was supported by the Faculty of Engineering, Chiang Mai University, and we also wish to thank the Supply Chain and Engineering Management Research Unit (SCEM), Chiang Mai University, for providing research facilities.

Conflicts of Interest: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Ahmed, Daiyaan, Ronhit Neema, Nishant Viswanadha, and Ramani Selvanambi. 2022. Analysis and Prediction of Healthcare Sector Stock Price Using Machine Learning Techniques: Healthcare Stock Analysis. *International Journal of Information System Modeling and Design (IJISMD)* 13: 1–15. [CrossRef]

Alnabulsi, Khalil, Emira Kozarević, and Abdelaziz Hakimi. 2023. Non-Performing Loans and Net Interest Margin in the MENA Region: Linear and Non-Linear Analyses. *International Journal of Financial Studies* 11: 64. [CrossRef]

Alzubi, Jafar, Anand Nayyar, and Akshi Kumar. 2018. Machine learning from theory to algorithms: An overview. *Journal of Physics: Conference Series* 1142: 012012. [CrossRef]

Anzanello, Michel Jose, and Flavio Sanson Fogliatto. 2011. Learning curve models and applications: Literature review and research directions. *International Journal of Industrial Ergonomics* 41: 573–83. [CrossRef]

Banegas, Ayelen, Gabriel Montes-Rojas, and Lucas Siga. 2022. The effects of US monetary policy shocks on mutual fund investing. *Journal of International Money and Finance* 123: 102676. [CrossRef]

Berry, William Dale, and Stanley Feldman. 1985. *Multiple Regression in Practice*. Newcastle upon Tyne: Sage.

Bolaman, Özge, and Pınar EVRİM. 2014. Effect of investor sentiment on stock markets. *Finansal Araştırmalar ve Çalışmalar Dergisi* 6: 51–64. [CrossRef]

Bolboacă, Roland, and Piroska Haller. 2023. Performance Analysis of Long Short-Term Memory Predictive Neural Networks on Time Series Data. *Mathematics* 11: 1432. [CrossRef]

Brogaard, Jonathan, and Abalfazl Zareei. 2023. Machine learning and the stock market. *Journal of Financial and Quantitative Analysis* 58: 1431–72. [CrossRef]

Chatterjee, Ananda, Hrisav Bhowmick, and Jaydip Sen. 2021. Stock price prediction using time series, econometric, machine learning, and deep learning models. Paper presented at the 2021 IEEE Mysore Sub Section International Conference (MysuruCon), Hassan, India, October 24–25.

Cheng, Leonardo, and Kartika Dewi. 2020. The effects of inflation, risk, and money supply on mutual funds performance. *Journal of Applied Finance and Accounting* 7: 29–34. [CrossRef]

Chimmula, Vinay Kumar Reddy, and Lei Zhang. 2020. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons & Fractals* 135: 109864.

Diebold, Francis X., and Roberto S. Mariano. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* 13: 253–63.

Dillender, Marcus, Andrew Friedson, Cong Gian, and Kosali Simon. 2021. Is healthcare employment resilient and "recession proof"? *INQUIRY: The Journal of Health Care Organization, Provision, and Financing* 58: 00469580211060260.

Ersin, Özgür Ömer, and Melike Bildirici. 2023. Financial Volatility Modeling with the GARCH-MIDAS-LSTM Approach: The Effects of Economic Expectations, Geopolitical Risks and Industrial Production during COVID-19. *Mathematics* 11: 1785. [CrossRef]

Gu, Shihao, Bryan Kelly, and Dacheng Xiu. 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies* 33: 2223–73. [CrossRef]

Gülmez, Burak. 2023. Stock price prediction with optimized deep LSTM network with artificial rabbits optimization algorithm. *Expert Systems with Applications* 227: 120346. [CrossRef]

Gyamfi Gyimah, Adjei, Bismark Addai, and George Kwasi Asamoah. 2021. Macroeconomic determinants of mutual funds performance in Ghana. *Cogent Economics & Finance* 9: 1913876.

Heshmaty, Behrooz, and Abraham Kandel. 1985. Fuzzy linear regression and its applications to forecasting in uncertain environment. *Fuzzy Sets and Systems* 15: 159–91. [CrossRef]

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9: 1735–80. [CrossRef] [PubMed]

Janiesch, Christian, Patrick Zschech, and Kai Heinrich. 2021. Machine learning and deep learning. *Electronic Markets* 31: 685–95. [CrossRef]

Jariyapan, Prapatchon, Jittima Singvejsakul, and Chukiat Chaiboonsri. 2022. A Machine Learning Model for Healthcare Stocks Forecasting in the US Stock Market during COVID-19 Period. *Journal of Physics: Conference Series* 2287: 012018. [CrossRef]

Jasra, Javed Mahmood, Rauf I Azam, and Muhammad Asif Khan. 2012. Impact of macroeconomic variables on stock prices: Industry level analysis. *Actual Problems of Economics* 134: 403–12.

Jolliffe, Ian T., and Jorge Cadima. 2016. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374: 20150202. [CrossRef]

Kang, Yushan, Jian Xing, and Shanhui Zhao. 2022. Influencing Factors of Investment for Companies. Paper presented at the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022), Harbin, China, January 21–23.

Kittichotsatsawat, Yotsaphat, Nakorn Tippayawong, and Korrakot Yaibuathet Tippayawong. 2022. Prediction of arabica coffee production using artificial neural network and multiple linear regression techniques. *Scientific Reports* 12: 14488. [CrossRef]

Kumar, P. Rajendra, and E. Bala Krishna Manash. 2019. Deep learning: A branch of machine learning. *Journal of Physics: Conference Series* 1228: 012045.

Li, Zheming. 2020. Economic Policy Uncertainty and Mutual Fund's Risk Adjusting Behavior in China. *Modern Economy* 11: 609–19. [CrossRef]

Lin, Eric C. 2018. The effect of Dow Jones industrial average index component changes on stock returns and trading volumes. *The International Journal of Business and Finance Research* 12: 81–92.

Mokhlis, Nur Hanis Mohd, Nur Anira Ahmad Burhan, Nur Fatin Ainsyah Roeslan, Siti Nur Aishah Zainal Moin, and Ummu Aiman Mohd Nur. 2021. Forecasting healthcare stock price using arima-garch model and its value at risk. *International Journal of Business and Economy* 3: 127–42.

Ouyang, Qi, Yongbo Lv, Jihui Ma, and Jing Li. 2020. An LSTM-based method considering history and real-time data for passenger flow prediction. *Applied Sciences* 10: 3788. [CrossRef]

Panigrahi, Ashok, Pradhum Karwa, and Pushkin Joshi. 2019. Impact of macroeconomic variables on the performance of mutual funds: A selective study. *Journal of Economic Policy & Research October* 15: 1–13.

Qureshi, Fiza, Ali M Kutan, Izlin Ismail, and Chan Sok Gee. 2017. Mutual funds and stock market volatility: An empirical analysis of Asian emerging markets. *Emerging Markets Review* 31: 176–92. [CrossRef]

Sarker, Iqbal H. 2021. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science* 2: 420. [CrossRef] [PubMed]

Sen, Jaydip, Sidra Mehtab, Abhishek Dutta, and Saikat Mondal. 2021. Precise stock price prediction for optimized portfolio design using an LSTM model. Paper presented at the 2021 19th OITS International Conference on Information Technology (OCIT), Bhubaneswar, India, December 16–18.

Shah, Jaimin, Darsh Vaidya, and Manan Shah. 2022. A comprehensive review on multiple hybrid deep learning approaches for stock prediction. *Intelligent Systems with Applications* 16: 200111. [CrossRef]

Slinker, Bryan K., and Stanton A Glantz. 1988. Multiple linear regression is a useful alternative to traditional analyses of variance. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 255: R353–R367. [CrossRef] [PubMed]

Sonkavde, Gaurang, Deepak Sudhakar Dharrao, Anupkumar M Bongale, Sarika T Deokate, Deepak Doreswamy, and Subraya Krishna Bhat. 2023. Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis and discussion of implications. *International Journal of Financial Studies* 11: 94. [CrossRef]

Subhani, Muhammad Imtiaz, Amber Osman, and Ameet Gul. 2010. *Relationship between Consumer Price Index (CPI) and KSE-100 Index Trading Volume in Pakistan and Finding the Endogeneity in the Involved Data*. MPRA Paper 26375. Available online: https://mpra.ub.uni-muenchen.de/29712/ (accessed on 1 November 2023).

Van Houdt, Greg, Carlos Mosquera, and Gonzalo Nápoles. 2020. A review on the long short-term memory model. *Artificial Intelligence Review* 53: 5929–55. [CrossRef]

Wanaset, Apinya. 2018. The relationship between capital market and economic growth in Thailand. *Journal of Economics and Management Strategy in Thailand* 5: 25–38.

Wong, Hock Tsen. 2022. The impact of real exchange rates on real stock prices. *Journal of Economics, Finance and Administrative Science* 27: 262–76. [CrossRef]

Zhou, Xianzheng, Hui Zhou, and Huaigang Long. 2023. Forecasting the equity premium: Do deep neural network models work? *Modern Finance* 1: 1–11. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Journal of
Theoretical and Applied
Electronic Commerce Research

MDPI

Article

The Impact of Academic Publications over the Last Decade on Historical Bitcoin Prices Using Generative Models

Adela Bâra [1] and Simona-Vasilica Oprea [2,*]

[1] Academy of Romanian Scientists, Ilfov 3, 050044 Bucharest, Romania; bara.adela@ie.ase.ro
[2] Department of Economic Informatics and Cybernetics, Bucharest University of Economic Studies, No. 6 Piaţa Romană, 010374 Bucharest, Romania
* Correspondence: simona.oprea@csie.ase.ro; Tel.: +40-752-29-44-22

Abstract: Since 2012, researchers have explored various factors influencing Bitcoin prices. Up until the end of July 2023, more than 9100 research papers on cryptocurrencies were published and indexed in the Web of Science Clarivate platform. The objective of this paper is to analyze the impact of publications on Bitcoin prices. This study aims to uncover significant themes within these research articles, focusing on cryptocurrencies in general and Bitcoin specifically. The research employs latent Dirichlet allocation to identify key topics from the unstructured abstracts. To determine the optimal number of topics, perplexity and topic coherence metrics are calculated. Additionally, the abstracts are processed using BERT-transformers and Word2Vec and their potential to predict Bitcoin prices is assessed. Based on the results, while the research helps in understanding cryptocurrencies, the potential of academic publications to influence Bitcoin prices is not significant, demonstrating a weak connection. In other words, the movements of Bitcoin prices are not influenced by the scientific writing in this specific field. The primary topics emerging from the analysis are the blockchain, market dynamics, transactions, pricing trends, network security, and the mining process. These findings suggest that future research should pay closer attention to issues like the energy demands and environmental impacts of mining, anti-money laundering measures, and behavioral aspects related to cryptocurrencies.

Keywords: cryptocurrency; research publication; topic modelling; latent Dirichlet allocation; sentiment analysis; Bitcoin prices

1. Introduction

The emergence of writing and money played a very important role in human history. They were tools that conveyed knowledge and intensified trade using a unified value that made it easier to exchange goods and services between people. The evolution of money probably began with clay inscriptions and continued with coined money, which are universal trading instruments built on people's trust. Blockchain and cryptocurrencies emerged after 2008 when Lehman Brothers went bankrupt, shocking financial markets and shaking confidence in banks. They represented, according to what Satoshi claimed, "the absolute autonomy of money" [1] and a response to the financial crisis. Digital technology's progress has shaped the performance and standards, imposing new payment methods and new money accepted by individuals, organizations and communities. Today, Bitcoin is the most popular virtual coin that has grabbed attention due to its spectacular price volatility. Out of the 22,900 crypto coins (with 8800 active), as of August 2023, Bitcoin's dominance had reached 49%. Most of its fluctuations were allegedly attributed to public opinion, influencers and social networks.

Bitcoin was born along with blockchain technology and while blockchain is perceived as beneficial to many fields, Bitcoin is highly controversial [2–4]. The high carbon footprint and energy consumption required by mining processes have led many researchers to call

it dirty [5–7]. It was estimated that Bitcoin mining is roughly responsible for 65.4 $MtCO_2$ yearly, which is comparable to Greece's emissions. Moreover, it was reported that the carbon intensity of Bitcoin mining increased by 17% in 2021 compared to 2020. In May 2021, specialized hardware devices (servers) worldwide consumed approximately 13 GW for mining processes. Inner Mongolia was one of the first Chinese provinces which banned crypto mining in March 2021 due to environmental concerns. Until June 2021, crypto mining bans were imposed in Sichuan and Xinjiang as well. In November 2021, Sweden also called for a ban on cryptocurrency mining as the usage of renewable energy to supply crypto mining would delay the energy transition towards a green environment [8]. Its price fluctuations have been remarkable, leading investors to rate it as a high-risk, less predictable investment tool. Security, privacy, money laundering and payments to the dark market are issues that have to be further investigated and tackled [9–12]. Aiming to find out the status quo of blockchain entrepreneurial innovations and research and how can these innovations contribute to the economic and sustainable development of society, these authors approached a combined method using web mining and extracting topics from unstructured data on the blockchain from Crunchbase, similar websites and research publications [13]. The analysis indicated that blockchain innovations have gone beyond the financial sector, addressing critical infrastructures such as the supply chain and energy delivery. More potential is envisioned in several promising application fields such as energy communities, renewable energy sources, digital twins, sustainable energy, voting, traceability or integrity verification. A blockchain solution was envisioned to solve the local energy trading aspects [14]. Furthermore, a conceptual architecture design of a blockchain solution for e-voting in elections at the university level was proposed in [15]. The power of a blockchain-based supply chain was further examined in [16].

Moreover, the way Bitcoin and other cryptocurrencies are perceived depends largely on the cultural perspective and ideological nuances at the regional or even at the country level. How cryptocurrencies are perceived in Japan and Sweden was analyzed in [17], examining attitudes and discourses on social media. These discourses focused on the role of blockchain technology in climate change, its environmental impact, the legitimate use of cryptocurrencies, politicians' tax evasion, payments to the dark markets or other ethical aspects. By using topic modelling in relation to the social media posts on cryptocurrencies, the authors studied the peoples' attitudes towards cryptocurrencies in these two countries, discerning advantages, opportunities and ethical challenges that were related to cryptocurrencies. The authors demonstrated that autonomy and money are well connected. They interpreted cryptocurrencies as a monetary innovation brought by society's development of enhanced freedom, trading and autonomy. Like in ancient Greece, cryptocurrencies can be a trading opportunity for over two billion unbanked poor people that are not part of the financial system. Trade would offer them an opportunity to overcome poverty and precarious living conditions. The results of the analysis showed that the perception of cryptocurrencies highly depends on the trust in public and private institutions. Although both Japan and Sweden have a high degree of trust in institutions, Sweden tends to value the principle of autonomy more compared to Japan, who seems to be more pragmatic, as shown in the online forums. In [18], critics argued that cryptocurrency mining wastes a lot of energy, while supporters claimed it is environmentally friendly. This study introduces mining domestic production (MDP) to evaluate Bitcoin mining's output and carbon emissions in China, comparing it to three traditional industries. The findings revealed that Bitcoin mining's energy efficiency is not always the worst. The paper offers a fresh viewpoint on assessing Bitcoin mining's profitability against carbon emissions, relative to other sectors. Additionally, it suggests that Bitcoin mining might help developing countries expand their electrical infrastructure and earn revenue. From the economic and sentiment analysis perspectives, several research papers are analyzed [19]. The differential influence of social media sentiment on cryptocurrency returns and price volatility during the COVID-19 pandemic were investigated in [20]. Whether cryptocurrencies provide a viable hedging mechanism for benchmark index investors was further studied in [21]. The

non-linear causal linkages of EPU and gold with major cryptocurrencies during bull and bear markets were examined [22], whereas herding behavior and price convergence clubs in cryptocurrencies during bull and bear markets were investigated in [23].

Although more than 9100 papers about Bitcoin and cryptocurrencies were written and indexed in Web of Science Clarivate platform, the relation between academic publications over the past 10 years and Bitcoin prices has yet to be examined. The scope of this paper is to reveal the main topics that concern researchers in relation to high-risk, less predictable investment instruments. By extracting the relevant topics, we aim to find gaps and potential new topics that have been less approached by the academic community. To identify topics, latent Dirichlet allocation (LDA) is employed. Furthermore, the document-topic density (α), word-topic density (β) and the optimal number of topics (K) are investigated. Equally interesting is the discovery of the sentiment of this community in relation to Bitcoin/cryptocurrencies and its effect on prices. Although sentiment analysis has been extensively performed on tweets, forums, other social media posts and reports [24], the sentiment of research publications has not been investigated yet. Additionally, it is intriguing to observe whether there are correlations between academic opinions and Bitcoin prices or connections between thousands of abstracts and prices.

This paper aims to answer the following three questions:

(RQ1) What are the relevant topics investigated by researchers in relation to cryptocurrencies and potential gaps?

(RQ2) What is the academic community's sentiment in relation to Bitcoin/cryptocurrencies?

(RQ3) Is there a connection between academic publications and Bitcoin's price evolution?

The contribution of this paper lies in its comprehensive approach to understanding the academic landscape surrounding Bitcoin and cryptocurrencies. The key contributions are as follows:

(a) An extensive review of cryptocurrency research and identification of key themes: The study analyses over 9100 research papers on cryptocurrencies, providing a broad overview of the academic focus in this field since 2012. This extensive review helps in identifying the predominant themes and trends in cryptocurrency research. Employing LDA to analyze unstructured abstracts, the study identifies key topics within the vast body of cryptocurrency literature. This approach offers a structured way to understand the main areas of academic interest and research. By calculating perplexity and topic coherence metrics, the research determines the optimal number of topics, adding rigor to the topic modelling process and ensuring the relevance and clarity of the topics identified;

(b) The use of advanced NLP techniques: The application of BERT-transformers and Word2Vec to process abstracts for predicting Bitcoin prices represents an innovative use of NLP in financial analysis;

(c) An investigation of the academic influence on Bitcoin prices: The study explores a relatively unexamined area—the impact of academic publications on Bitcoin prices. The findings indicate a weak connection, providing insights into the influence of academic research on market dynamics;

(d) The sentiment analysis of academic publications: Unlike prior studies that focused on the sentiment analysis of social media, this study extends the analysis to academic writings, offering a new perspective on how academic sentiment correlates with Bitcoin prices;

(e) Highlighting research gaps and future directions: By extracting relevant topics and sentiments, the study identifies gaps and less-explored areas in cryptocurrency research, suggesting future directions, such as the impact of mining on energy and the environment, anti-money laundering measures, and behavioral aspects of cryptocurrency use. Overall, this paper makes a significant contribution by blending topic modelling, sentiment analysis, and financial prediction techniques to provide a multifaceted view of the academic discourse on cryptocurrencies and its potential impact on market behavior.

The novelty in this paper lies in several key areas: (1) The unexplored relationship between academic publications and Bitcoin prices: While there has been extensive research on various factors influencing Bitcoin prices, this study is novel in its examination of the direct relationship between academic publications and Bitcoin price movements. It explores whether scientific writing in this domain has any significant impact on the market; (2) The use of advanced analytical NLP techniques: The employment of LDA to analyze unstructured abstracts from over 9100 research papers is a novel approach. This method, combined with perplexity and topic coherence metrics, is used to determine the optimal number of topics; (3) BERT-transformers and Word2Vec for price prediction: The study innovatively uses BERT-transformers and Word2Vec to vectorize (process) abstracts and assess their potential in predicting Bitcoin prices. This application of advanced NLP technologies in the financial analysis of cryptocurrencies represents a novel approach; (4) The sentiment analysis of the academic community: While sentiment analysis has been widely conducted on social media platforms, this study breaks new ground by focusing on the sentiment within academic research publications and its potential effect on Bitcoin prices; (5) The identification of research gaps and emerging topics: By extracting and analyzing topics from a vast number of research papers, the study identifies gaps and new topics that have been less explored by the academic community, offering directions for future research; (6) A comprehensive scope of analysis: The paper's aim to answer three research questions about relevant topics, academic sentiment, and the connection between research and Bitcoin prices covers a comprehensive scope that has not been collectively explored before in the context of cryptocurrency research. Thus, the novelty of this study is its comprehensive and multidimensional analysis of the relationship between academic research and the Bitcoin market.

The current research is structured in five sections: in the Section 1 a brief introduction regarding our purpose, motivation and research questions is presented. Furthermore, we expose three research questions that will be answered in the Section 5. In Section 2, the most relevant previous publications are summarized indicating the main findings. The research methodology is described in Section 3, and it is focused on the LDA technique implementation to extract the relevant topics from the abstracts. Results are presented in Section 4 and conclusions in Section 5.

2. Literature Review

One of the input datasets included in this paper consisted of the list of 9105 publications on Bitcoin and cryptocurrency extracted from the Web of Science Clarivate platform (https://clarivate.com/, accessed 30 July 2023). Therefore, we exhaustively searched the most recent publications related to the topic modelling, sentiment analysis, research publications, academic community opinion and prediction related to the cryptocurrencies in general and Bitcoin in particular. Extracting news from media, social posts and even research content and the implementation of natural language processing (NLP) techniques have been extensively investigated over the past decade. Moreover, topic modelling was one of the NLP applications studied in several papers.

2.1. Topic Modelling

The relation between the international news and Bitcoin prices was analyzed in [25]. The news released from 2018 and 2020 were investigated using an LDA-based topic modelling approach. The text of 4218 cryptocurrency-related news articles from 60 countries were searched for significant topics, identifying 18 relevant topics related to security, financial issues, the market and economy. The results indicated that the news led to an impact of their content on the Bitcoin price variations' increasing volatility. After the publication, there was a negative impact on the next 24 h of prices.

Blockchain technology has been evolved together with cryptocurrencies and has found applicability in various fields [26]. It is one of the promoters of cryptocurrencies. Therefore, numerous research papers focused on investigating this technology. More

than 900 blockchain-related papers published by the IEEE, Springer and ACM publishing houses along with other databases were selected and their texts formed a corpus that was analyzed using LDA. The scope was to identify topics that require more investigation. The authors found 15 research directions or gaps that can be analyzed in relation to blockchain technology. Warehouse receipts are key in supply chain finance [27]. Traditional pledging methods are inefficient and insecure, leading to issues like duplicate pledges. To combat this, the authors introduce a Blockchain-based Digital Asset Platform (BDAP) with enhanced security and multi-party certification. Testing shows the BDAP's efficiency with an average response time of 1.441 s under a high user load. However, its adoption is slow, with limited bank participation, suggesting a gradual change in the industry's traditional mindset. Ref. [28] analyzed blockchain technology through six core layers: its application, contract, actuator, consensus, network, and data. It reviews the literature to understand the blockchain's global applications, finding China's focus on practical applications in industries and smart cities. In contrast, international research focuses on the blockchain in finance, integrating crypto assets with traditional industries like payment systems. The findings highlight the interplay between smart cities and cryptocurrencies, suggesting a mutually reinforcing relationship. Other applications of LDA and sentiment analysis were also found [29,30], revealing a broad spectrum of implementation.

Data from firms' reporting files are examined in [31], focusing on the blockchain technologies, cryptocurrencies and related applications. Both disclosers about blockchain and cryptocurrencies were analyzed using text-processing techniques. LDA was applied to extract relevant topics from the two types of disclosures. Five topics were underlined by the LDA: solutions related to blockchain implementation, factors that may bring risks, business models, and market-related services: payment and transactions. Furthermore, the two types of disclosures were analyzed from the relevance point of view and the authors found out that the blockchain solution and risk issues brought a positive value relevance, whereas disclosures regarding Bitcoin transactions and other cryptocurrencies brought a negative value relevance. Therefore, the firms positively valued blockchain applications, whereas cryptocurrencies and their related issues became less appreciated.

The literature related to blockchain technology published between 2013 and 2018 was investigated in [32], aiming to understand the status of blockchain research and plan the next research steps. An LDA was applied, revealing that most of the research was in the computer science and business research area. The main topics of the current research are related to technology architecture, financial applicability, privacy and security issues and smart applications. Moreover, the research envisioned three next stages: technology, business application and integration with IoT, AI and other technologies. Additionally, the authors discovered two strong connections between the blockchain and cryptocurrencies and between the blockchain and various applications. The software developers' posts from two blockchain-related Stack Exchange sites were investigated with an LDA to identify relevant topics [33] in an attempt to find the challenges that software developers encounter. The results showed that the posts regarding the blockchain increased in occurrence, while those regarding mining cryptocurrencies decreased. The authors considered that more documentation on blockchain development would support developers' communities that may lack supporting materials and guidelines. Moreover, there is an increasing interest in blockchain technology, spurred by Bitcoin's success and its expanding range of applications [34]. More studies emphasized the need for high-performance data interaction within blockchain systems. An overview of blockchain's development is provided, examining research efforts to enhance blockchain performance from on-chain, off-chain, and cross-chain interaction technology perspectives. Special attention is given to cross-chain interaction technologies, with the results showing that cross-chain technologies significantly enhanced blockchain performance.

The sentiment analysis, Bitcoin price and topic modelling are interconnected. Sentiment analysis provides insight into the emotional tone of discussions surrounding Bitcoin. Positive sentiment can attract more buyers and investors, potentially leading to price in-

creases, while negative sentiment can lead to price drops. Topic modelling helps identify the key subjects being discussed in relation to Bitcoin. By understanding the topics driving sentiment, investors can anticipate potential market-moving events and reactions. Both sentiment analysis and topic modelling contribute to the broader understanding of market dynamics, which can be leveraged by traders, investors, and analysts to make informed decisions about Bitcoin. It is important to note that while these methods can provide valuable insights, the cryptocurrency market is complex and influenced by a wide range of factors. Therefore, a comprehensive approach to analysis, including technical, fundamental, and sentiment-based insights, is often used to make informed decisions in this volatile market.

2.2. Sentiment Analysis

Twitter sentiment was analyzed to find whether the price of Bitcoin would increase or decrease and its magnitude [35]. The authors examined emotions conveyed by the tweets and their volume. The challenge was to set the time window in which the emotion is transmitted and impacts the price and, thus, becomes a real predictor. The study managed to predict the Bitcoin price direction and magnitude with good accuracy. Moreover, the NFTs were monitored, and the prediction was extended to the NFTs, namely CryptoPunks, which is one of the most well-known collections on the market. Other authors investigated Elon Musk's activity on Twitter in January 2021 and its impact on Bitcoin prices [36]. The analysis involved sentiment analysis on tweets after January 2021 and data extracted from Binance to understand the relationship between tweets and price evolution. The results showed that the volume of tweets did significantly increase. The volume of tweets was strongly and positively correlated with Bitcoin's price evolution. However, the effect of the tweets' sentiment on price proved to be weak; therefore, sentiment was not included as a credible predictor. Moreover, the study aimed to understand the influence of social media influencers that may affect Bitcoin's evolution and have a significant impact on its price. A similar study analyzed the effects of the COVID-19 pandemic and public sentiment on the cryptocurrencies' prices. The authors implemented the exponential generalized autoregressive conditional heteroskedasticity model, a quantile regression and a sentiment analysis showing that cryptocurrencies were less predictable during the COVID-19 pandemic [37].

A light gradient boosting machine (LGBM) classifier was employed in [38]. The results showed that 78.06% and 94.03% of hourly and daily bullish market movements can be attributed to public tweets, whereas 83.08% and 94.60% of hourly and daily bearish market movements could be justified by public tweets. The authors examined the interval from September 2017 to September 2022, analyzing more than 28 million tweets containing Bitcoin as a keyword. Machine learning (ML) and NLP techniques were applied to vectorize text, calculate semantic distances, extract keywords and encode variables. Furthermore, ML and NLP were employed to extract insights from tweets and expert ratings obtaining the best results of assessing the sentiment with a support vector machine (SVM) classifier. The authors used 68,281 tweets from 57 initial coin offerings (ICO) from 4 industries (the cryptocurrency, platform, business services, and entertainment) [39].

The Bitcoin price direction was predicted using a linear discriminant analysis-based classifier and sentiment analysis [40]. Price information and crypto-related news headlines were collected to forecast the day-ahead price direction. The SVM outperformed the other forecasting algorithms. However, all of them show better accuracy when predicting an increase in the day-ahead price compared to forecasting a decrease in the price. The inclusion of news sentiment increased the accuracy by 0.585. News and information from social groups in financial markets were investigated in [41]. The authors analyzed the relationships between moods and price time series to predict prices. The sentiment moods were calculated using the probability distribution of the news and were vectorized using a BERT-based transformer language model. Recurrent neural networks were employed to forecast the prices. Additionally, 15,000 tweets on cryptocurrency were analyzed from the emotional point of view evaluating the emotion score for anger, anticipation, disgust, fear,

joy, sadness, surprise, trust and the two basic sentiments: negative and positive. A total of 53,077 sentiments were extracted from the 15,000 tweets. The results showed that the data sample consisted in positive sentiments that might lead to an increase in investment in the decentralized finance market. This study indicated how ML algorithms measured the emotions from tweets to understand the implications of cryptocurrencies' prices [42].

A few months of Twitter data before and after the COVID-19 pandemic were analyzed in [43], applying the latent semantic analysis and decomposition of singular values. Its scope was to analyze the public's opinion on Bitcoin and cryptocurrencies ex-ante and ex-post the COVID-19 pandemic and understand the relevant topics that led to negative emotions related to cryptocurrencies and trading. The aim was to underline and share the findings (topics). Furthermore, an analysis of the change in discussions on social media related to the Bitcoin price was presented in [44]. The authors proposed a Word2Vec vectorized topic model that identifies which topics on social networks had an impact on Bitcoin prices from 2017 to 2018 which were characterized as a higher volatile interval. The authors tested the words' change in frequency within four intervals and compared four word2vec models (combining a continuous bag of words, cbow, skip-gram, sg, Hierarchical Softmax, HS, and Negative Sampling, NEG) to evaluate their consistency and performance. Eight topics were identified when prices shifted to a decreasing slope: wallet, fork, transfer, posts, block size, confirmation, exchanges, and password; whereas three to five topics were identified when prices shifted to an increasing slope: startup, East Asia, ICO, competition, and lighting network.

Cryptocurrency ecosystems and social media environments were investigated in [45]. A Hawkes' model was employed together with other NLP techniques to conceive an empirical analysis to figure out the relationship between cryptocurrencies' prices and social media from January 2019 onwards, concentrating on the rise of the Ethereum and Bitcoin. The relationship between the shifts in cryptocurrencies' prices, sentiment and topic conversations on social media was investigated using the Hawkes' model. The results showed that some topics and sentiment in social media forego certain price movements. Topics such as trading, governments and exchange currency (Ethereum cryptocurrency) negatively impacted Ethereum and Bitcoin prices. Discussions related to investments led to price increases, whereas discussions related to decentralized and technological applications led to price decreases. Furthermore, 14.156 posts from the Reddit blockchain platform were analyzed and topics such as currency inflation, smart contracts auditing, crypto comparison (Bitcoin, Ethereum, Hyperledger Fabric, Cardano, and the comparison between Stellar and Ripple proved to be more frequent than others) and blockchain trends (which wallet is the best for storing tokens, most arguments concerning Stellar and Ethereum trends) were revealed in [46]. The authors explored the users' topics extracted from conversations about blockchain trends, technologies, platforms, smart contract development and auditing using topic modelling and the BERT-transformer model.

The influence on Bitcoin prices from popular social networks makes prices extremely volatile and speculative. However, the influences of users are not equal and a hypertext-induced topic selection algorithm was used to split the dataset into two groups based on the users' influence [47]. Topic modelling was employed to extract topics from the discourses of the two groups. Significant differences in opinions were identified. Additionally, the results indicated that the opinions of the leaders were not in line with the majority. Moreover, in research, there is an inertia between the emergence of technology and publications [48]. However, substantial research has been undertaken regarding blockchain and Bitcoin. Their potential development and impact as well as their positive roles have to be enhanced. Solutions and cross-disciplinary exchanges are also required when it comes to energy consumption and the carbon footprint. The next steps in terms of research concepts, applications and future exploration were envisioned. Analyzing the sentiments of users and extracting the topics from social media via large data streaming are essential to making timely decisions. The model proposed in [49] provided a dynamic and scalable topic modelling over data streaming and a sentiment analysis at the topic level. Posts related to

Ethereum, Bitcoin and Facebook from Twitter were investigated, and the results offered a large data-scale topic model on social media streaming and a sentiment analysis at the topic level.

2.3. Crypto Price Prediction

To explore the extensive scope of analysis in predicting Bitcoin prices is one of our objectives. Researchers considered a variety of factors to forecast Bitcoin prices with maximum accuracy. The year 2021 witnessed significant volatility, with rapid transitions between bull and bear markets. This period saw both substantial losses and gains, prompting researchers to identify the key variables and indices that were most influential on Bitcoin price movements. Contrary to predictions that the fluctuations seen in 2021 would not recur, by the end of 2023, there was a sudden surge in Bitcoin prices, indicating the onset of a new bull market. However, Bitcoin is still considered a high-risk investment instrument whose price is difficult to predict. Using long short-term memory (LSTM), 47 input variables for 10 months were trained to predict the Bitcoin price movements resulting in an error of 3.52% [50]. Considering the transaction data, namely the buy and sell orders from Coinbase, which is one of the major digital currency exchange platforms, the investors' sentiment was measured. Then, using a bootstrapped quantile regression approach, the authors mentioned a relevant relation between sentiment and Bitcoin prices [51]. Additionally, day-ahead Bitcoin price predictions using random forest regression and LSTM, including variables which have impact on the Bitcoin price prediction, were depicted in [52]. A three-year dataset was considered (from 2015 to 2018), three stock market indexes (NASDAQ, S&P500, DJI) from the US, oil prices, CO_2 certificates and ETH prices were proven to impact the Bitcoin prices. However, from 2018 onwards, JP225, which is a Japanese stock market index, and ETH prices increased in importance. The accuracy of the prediction model depended on the interval and the number of lags. The model with one lag provided the best accuracy. The relation between Bitcoin and US stock markets was further investigated in [53], estimating the Bitcoin capacity to forecast the US composite and sectoral stock indices using daily data from November 2017 to December 2021. The results showed that the Bitcoin price is a significant predictor for the US stock market volatility. The findings also indicated an inverse relation between Bitcoin prices and the US stock prices emphasizing the importance of tracking Bitcoin prices for both practitioners and policy makers.

The effects of the interest rate on the Bitcoin price were investigated in [54] using a structural vector autoregressive (SVAR) model. The input dataset was extracted from January 2012 to October 2022 considering the following variables: VIX, which is the Chicago Board Options Exchange's (CBOE) Volatility Index—a popular measure of the stock market expectation of volatility that is based on S&P500 index options, the interest rate spread, positive real interest rate and negative real interest rate, gold price, oil price and DXY, which is the U.S. Dollar Index that tracks the strength of the dollar against six currencies. DXY goes up when the U.S. dollar gains strength compared to other currencies. Based on the variance decomposition, the negative real interest rate shocks had a stronger impact than the positive real interest rate on the Bitcoin price. A long-run Bitcoin price prediction using regressive methods and ML methods, namely SVM, was investigated in [55] showing the Bitcoin trend to stock-to-flow and providing insights into Bitcoin's scarcity/abundance by comparing its current stock with the flow of new stock entering the market over a certain period of time and potential price trends. Both these approaches are data-driven. Addressing the challenge of identifying one-time change addresses in Bitcoin address clustering, an emerging issue in social computing, ref. [56] introduced an innovative clustering method. Traditional research in this area, often limited to specific transaction types, struggles with low recognition rates and high false positives. Their method, utilizing multi-conditional recognition, was tested with on-chain Bitcoin transaction data. It outperforms existing heuristics, identifying at least 12.3% more one-time change addresses.

As Bitcoin is one of the means of financing terrorism, the authors of [57] explored the Bitcoin prediction from an terrorism-based event perspective. They used a vector autore-

gressive (VAR) model to examine the financing capacity of Bitcoin and risk hedging. The results showed that terrorist attacks/incidents and brutality have a significant impact on Bitcoin price. These aspects indicated that the investors were concerned about the number of casualties and the frequency of the attacks. The effect of Bitcoin price on terrorism was also analyzed, but the influence was not significant. Econometrics were further employed to predict Bitcoin volatility using two-component CARR, GARCH and CGARCH models [58]. To estimate the effectiveness of the investors' sentiment in predicting the prices, the authors of [59] used the Bitcoin Misery Index (BMI) that results from trading cryptocurrencies and eliminating individuals judgements. The BMI indicates buy and sell opportunities for Bitcoin investors. When the BMI falls below 27, it indicates a strong buy signal. The higher the BMI is, the more likely the Bitcoin price may fall. Furthermore, they employed a bagged support vector regression to predict the Bitcoin prices. The study spanned from March 2018 to May 2022 and the results indicated that the index improved the accuracy of the prediction model. Additionally, a feature selection method enhanced the results of the prediction model providing the prices for the next 30 days. The relation between the Bitcoin price volatility and market concentration was described in [60], revealing the supply side of the Bitcoin ecosystem. It was demonstrated that the bounded market concentration in pooled mining capped the Bitcoin price variation obtaining a robust prediction model.

Moreover, the researchers investigated the connectedness between Bitcoin price and the energy consumption for mining processes that may have an impact on Bitcoin prices [61]. The input dataset was extracted from January 2014 to July 2021 and the authors used an artificial neural network (ANN) to predict the Bitcoin prices taking into account the energy consumption. The study showed that there is a strong correlation between the prices and the consumed energy. ANNs are also used in [62] to predict the prices of Bitcoin. The dataset spanned from January 2018 to September 2018 and was extracted from coindesk. A combination of price-related and lagged features was created to improve the prediction capabilities of ANNs. The results indicated that the Fitnet network with a trainlm Matlab function and 30 hidden neurons outperformed the other ANNs. In [63], the unstructured data from financial news and deep learning (LSTM) were employed to predict the Bitcoin prices. This method outperformed the combinations between other ML algorithms and datasets that did not include financial news. The Facebook prophet model was proposed in [64] to predict the prices of cryptocurrencies. The results were compared to LSTM and ARIMA. The FB prophet proved to outperform LSTM and ARIMA models as these two models have drawbacks that make them incompatible with a performant cryptocurrency price prediction. In [65], the Bitcoin price was used as a predictor for the currency exchange rates prediction. The proposed prediction model consisted of an autoregressive distributed lag (ADL) providing good results for daily horizons.

The emergence of Bitcoin futures contracts, the COVID-19 pandemic and their impact on the Bitcoin price were analyzed in [66]. The results indicated that Bitcoin's future contracts positively impacted its returns, but it had no impact on price volatility. According to [66], the COVID-19 pandemic did not influence Bitcoin returns nor volatility. However, strong movements took place between Bitcoin prices and COVID-19-related events (such as lockdowns, viruses' evolution, the number of dead, vaccination stages, etc.). As numerous research studies struggled to identify the variables that significantly contribute to Bitcoin or other crypto prices and sometimes the results remained inconsistent, the authors of [67] analyzed a wide variety of variables that were considered in previous research using an extreme bounds analysis (EBA), which is a large-scale sensitivity analysis capable of analyzing model uncertainty aspects. The results obtained using EBA indicated that market supply and demand, policy uncertainty and public interest are robust features capable of explaining the price fluctuations. Furthermore, global macroeconomic and financial events may explain Bitcoin price movements.

The text synthesizes research findings across three conceptual frameworks central to understanding Bitcoin and cryptocurrency: topic modeling, sentiment analysis and price

prediction, drawing on a variety of data sources including academic publications, social media, and news articles.

In topic modeling, the aim is to decipher the thematic structure within a vast corpus of text data related to Bitcoin and cryptocurrencies. Researchers employ methods like LDA to parse content from diverse sources, uncovering prevalent topics such as blockchain technology's applications, financial issues, market dynamics, and security concerns. This approach reveals the broad spectrum of themes discussed in the context of a blockchain and cryptocurrency, pointing out specific areas that require further research and the impact of news and social media on market perceptions.

Sentiment analysis seeks to measure the emotional tone and public sentiment towards Bitcoin and cryptocurrencies based on online discourse. By analyzing social media posts, news headlines and other text-based data, researchers assess how sentiment influences Bitcoin prices. The application of learning techniques in sentiment analysis shows a clear link between public sentiment and Bitcoin price fluctuations, demonstrating how positive or negative sentiments can significantly affect market movements. The analysis extends to examining the impact of influential figures and global events, such as the COVID-19 pandemic, on market sentiment and cryptocurrency prices.

Price prediction focuses on forecasting Bitcoin price movements through various analytical methods, including statistical models, machine learning algorithms and deep learning frameworks. These models incorporate a wide range of variables, from market indices and sentiment data to transaction volumes and global economic indicators, striving to predict price changes with accuracy. Despite the challenges posed by the volatile and speculative nature of Bitcoin, some models have achieved notable success in forecasting price movements. Research in this area also explores the role of external factors, such as global events and technological advancements, in influencing Bitcoin prices.

While significant strides have been made in predictive analytics, the unpredictable nature of Bitcoin prices highlights the ongoing need for exploring new data sources, analytical techniques and theoretical frameworks. This continuous exploration aims to improve prediction accuracy and identify factors driving Bitcoin and other cryptocurrencies.

As noticed in the literature review, although the three subsections approach distinct themes, such as topic modelling, sentiment analysis and price prediction, they often mingle, and numerous researchers combine them to extract valuable insights for investors. As mentioned above, Bitcoin and cryptocurrencies are in the limelight of research and grab huge interest from both researchers and industry. However, one can notice that numerous studies were concerned about sentiment analysis, price prediction and its determinants and only a few studies were focused on LDA and top modelling approaches. To the best of our knowledge, no previous study has included unstructured data from research publication and its relationship with Bitcoin prices.

3. Methodology

The link between sentiment analysis, Bitcoin price, and topic modelling lies in their potential to provide insights into the dynamics of the cryptocurrency market and the factors that underly it. Sentiment analysis involves analyzing textual data, such as news articles, social media posts, and online discussions, to determine the overall sentiment or emotional tone of the content. In the context of the cryptocurrency market, sentiment analysis can help gauge market participants' attitudes, opinions, and emotions about Bitcoin and other cryptocurrencies. Positive sentiment might be associated with optimism and a belief in price appreciation, while negative sentiment could indicate concerns and potential price declines. By monitoring sentiment, traders and investors can gain insights into market sentiment shifts that might impact Bitcoin prices. The price of Bitcoin is determined by supply and demand dynamics within the market. It can be further influenced by a wide range of factors, including market sentiment, potential regulatory developments, macroeconomic trends, technological advancements, and more. Sentiment analysis can play a role in understanding how emotional factors can impact buying and selling decisions,

thereby affecting Bitcoin price movements. Positive sentiment might lead to increased demand and rising prices, while negative sentiment could lead to selling pressure and price drops.

In order to perform a price prediction, the input data in text format is processed using BERTs (Bidirectional Encoder Representations from Transformers) and Word2Vec that involve distinct processes due to their different architectures and purposes in NLP. A BERT is first pre-trained on a large corpus of text using tasks such as masked language modelling (MLM) and next sentence prediction (NSP). In MLM, random words in a sentence are masked for the model to predict, while in NSP, a BERT learns to predict if one sentence logically follows another. After pre-training, the BERT is fine-tuned for specific tasks like sentiment analysis or question answering. This involves training on a smaller, task-specific dataset, updating the entire model's weights but much quicker than pre-training. For inference, the input text is tokenized, and special tokens are added [68,69]. The BERT then generates a contextual representation of each token, useful for various downstream tasks. Word2Vec, on the other hand, can be trained using either the continuous bag of words (CBOW) model or the skip-gram model. CBOW predicts a word based on its context, while skip-gram does the opposite. Once trained, Word2Vec provides a vector for each word, capturing semantic meanings and relationships. These vectors can be used for text similarity, clustering, or as features in machine learning models [70,71]. Unlike the BERT, Word2Vec only provides word-level embeddings and does not consider a sentence-level context. The BERT and Word2Vec differ significantly. The BERT provides context-sensitive embeddings where the representation of a word changes based on its context, whereas Word2Vec offers static embeddings. The BERT is more complex and resource-intensive compared to the simpler and less computationally demanding Word2Vec. In terms of application, the BERT is versatile for a range of NLP tasks, especially those requiring an understanding of context or sentence structure, while Word2Vec is more limited to word-level analysis. The BERT and Word2Vec are both potent in the field of NLP and are suited for different tasks and challenges in language processing.

Topic modelling is a technique used to extract underlying topics or themes from a collection of documents. In the context of cryptocurrency, topic modelling can help identify the key subjects or trends that are being discussed in news articles, social media posts, and online forums. By analyzing these topics, traders and investors can gain a better understanding of the factors driving sentiment and influencing Bitcoin price movements. For example, topics related to regulatory news, crypto-crime events, environmental impacts, technological advancements, market adoption, ICO, or macroeconomic events can all impact sentiment and subsequently influence Bitcoin price.

Revealing relevant topics from large texts can be a complex task. On one hand, they show the mainstream, but on the other hand they also indicate the gaps in the research field. Latent Dirichlet allocation (LDA) helps to explain a large text dataset by identifying the unobserved groups or topics that are also known as latent. In the text, words form a large bag of words in which each word is allocated to a topic. LDA is applied to gain insights into the realm of unstructured data. Initially, the LDA was proposed in genetics by Pritchard et al. [72] and then in machine learning by David Blei, Andrew Ng and Michael Jordan in the *Journal of Machine Learning Research* in 2003 [73]. It was also refined and described by Jelodar et al. Research papers (from 2003 to 2016) related to topic modelling based on LDA were investigated to discover the research progress, trends, challenges and applications in various sciences [74].

LDA is a probabilistic technique that aims at identifying the main topics in large collections of text data. When LDA is employed, probabilities are calculated using the Bayesian method and the expectation maximization (EM) algorithm that is based on iterations in order to find the maximum likelihood. The EM performs (1) an expectation step (known as E) which generates a function L for the expectation of the log-likelihood and (2) a maximization step (known as M) which computes the parameters ω that maximize the expected log-likelihood identified at the E step. These parameters are calculated to identify

the distribution of the latent variables. Having a set of observed variables X, a set of topics K and the model parameters ω, the likelihood function L, the maximum likelihood estimate (MLE) is calculated by maximizing the marginal likelihood of X:

$$L(\omega; X) = p(X|\omega) = \int p(X|K, \omega)p(K|\omega)dK \tag{1}$$

where X—observed variables;

K—the latent variables (topics);

ω—the model parameters;

L—the likelihood function.

The EM algorithm iteratively alternates the two steps (E and M). In the expectation step, the expected value $Q\left(\omega\middle|\omega^{(t)}\right)$ is defined:

$$Q\left(\omega\middle|\omega^{(t)}\right) = E_{K\sim p(\cdot|X,\omega^{(t)})}[\log p(X, K|\omega)] \tag{2}$$

In the M step, the parameters ω that maximize the expected values are found:

$$\omega^{(t+1)} = \underset{\omega}{\operatorname{argmax}} Q\left(\omega\middle|\omega^{(t)}\right) \tag{3}$$

$$\omega^{(t+1)} = \underset{\omega}{\operatorname{argmax}} E_{K\sim p(\cdot|X,\omega^{(t)})}[\log p(X, K|\omega)] \tag{4}$$

LDA identifies the latent topics in the text similar to the exploratory factor analysis (EFA) applied to numeric features. Moreover, a confirmatory factor analysis (CFA) takes latent variables and estimates whether the latent features via a measurement model explain the observed variables [75,76]. The LDA derives from the probabilistic latent semantic analysis (pLSA). While both methods are similar and require the user to specify the number of topics (similar to the K-means clustering method), LDA compared to pLSA offers a better disambiguation of words and a more accurate allocation of documents to latent topics.

The total probability of the LDA model is described in Equation (5):

$$P(W, Z, \omega, \varphi; \alpha, \beta) = \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{j=1}^{A} P(\omega_j; \alpha) \prod_{t=1}^{N} P\left(Z_{j,t}\middle|\omega_j\right) P\left(W_{j,t}\middle|\varphi Z_{j,t}\right) \tag{5}$$

A—the number of abstracts;

N—the number of words in a given abstract j;

K—the number of topics;

α—the parameter of the Dirichlet per-document topic distributions;

β—the parameter of the Dirichlet per-topic word distribution;

ω—the θ i topic distribution for abstract j;

φ—the φ k word distribution for topic i;

Z—the z i j—topic for the word t in abstract j;

W–a w i j—specific word.

The two types of analyses (the relevant topic extraction using LDA and the Bitcoin price prediction) are presented below following the data collecting, pre-processing, processing, analyzing and interpreting stages. The process of obtaining the relevant topics is performed in seven steps:

I. Topic extraction using Latent Dirichlet Allocation

Step 1: Input dataset WOS as dataframe1 (df1). Details regarding the data source are provided in Section 4;

Step 2: Pre-process the Abstract column from WOS (eliminating punctation, lowercasing characters, removing stopwords, etc., using a pipeline);

Step 3: Perform an exploratory data analysis (EDA) on the pre-processed Abstract—WordCloud;

Step 4: Create the dictionary and obtain the corpus that is a term document frequency;

Step 5: Set the number of topics, building the LDA model using the corpus and the number of topics;

Step 6: Calculate the coherence score and perplexity to find the optimal number of topics;

Step 7: Visualizing the topic using the inter-topic distance map and top 30 most salient terms.

II. Bitcoin price prediction

Step 1: Input dataset WOS as dataframe1 (df1), input dataset Bitcoin prices as dataframe2 (df2). Details regarding the data source are provided in Section 4;

Step 2: Merge df1 and df2 using the Month and Year columns;

Step 3: Pre-process the Abstract column from WOS (eliminating punctation, lowercasing characters, removing stopwords, etc., using a pipeline);

Step 4: Use TextBlob to analyze sentiment. Polarity is obtained from the pre-processed Abstract;

Step 5: Calculate and graphically visualize the percentages of positive, neutral and negative sentiments;

Step 6: The pre-processed Abstract column is vectorized using the BERT-transformer and Word2Vec (vector_size = 100, window = 5, workers = 4, sg = 0), using the skip-gram algorithm;

Step 7: Apply several regressors (random forest, histogram gradient boosting, eXtreme Gradient Boosting, light gradient boosting, linear regression, and a voting regressor) to predict the Bitcoin price. Test the result for different time horizons. Calculate performance metrics (MAE, RMSE, MAPE) and graphically visualize the results.

Thus, in this paper, we process text data using a range of NLP techniques to draw out pertinent insights. To transform the text into numerical features, we employ advanced vectorization methods such as the BERT and Word2Vec. These methods enable us to generate numerical vectors, facilitating the training of various machine learning models, including random forest, histogram gradient boosting, eXtreme Gradient Boosting, light gradient boosting, linear regression, and a voting regressor. While there are alternative vectorization techniques available, such as TF-IDF (term frequency–inverse document frequency), one-hot encoding or n-grams, we found them to be less efficient and less suitable for handling large data volumes. Additionally, we experiment with different configurations of the LDA to achieve optimal outcomes.

4. Results

One of the datasets was extracted from the Web of Science Clarivate platform on 30 July 2023. Two words were included in the search: Bitcoin and cryptocurrency. The total number of research papers that met the searching criteria summed up to 9105 and the number of columns totalized 72 (such as: Publication Type, Journal Abbreviation, Publisher, Authors, Affiliations, Abstract, Keywords, Article Title, Cited Reference Count, Times Cited, WoS Core, Times Cited, All Databases, ISSN, eISSN, Publication Date, Volume, Issue, DOI, DOI Link, Number of Pages, WoS Categories, Web of science Index, Research Areas, Highly Cited Status, UT (Unique WOS ID), Year, etc.).

The abstract was analyzed and converted from text to features. A total of 357 articles were excluded from the list as their abstracts did not exist. Moreover, 3839 articles were without a publication date that was necessary to be converted into a valid month. Therefore, a random function between one and twelve was applied. A total of 2361 papers were in Computer Science Information Systems, 2001 in Computer Science Theory Methods, and 1568 in Business Finance. Researchers from the USA wrote 1800 papers, 1647 were from China and 818 were from England. A total of 5704 papers were in the articles category, 2911 were proceeding papers and 309 were in the early access category. The number of publications related to Bitcoin and other cryptocurrencies progressively increased over time (as in Figure 1).

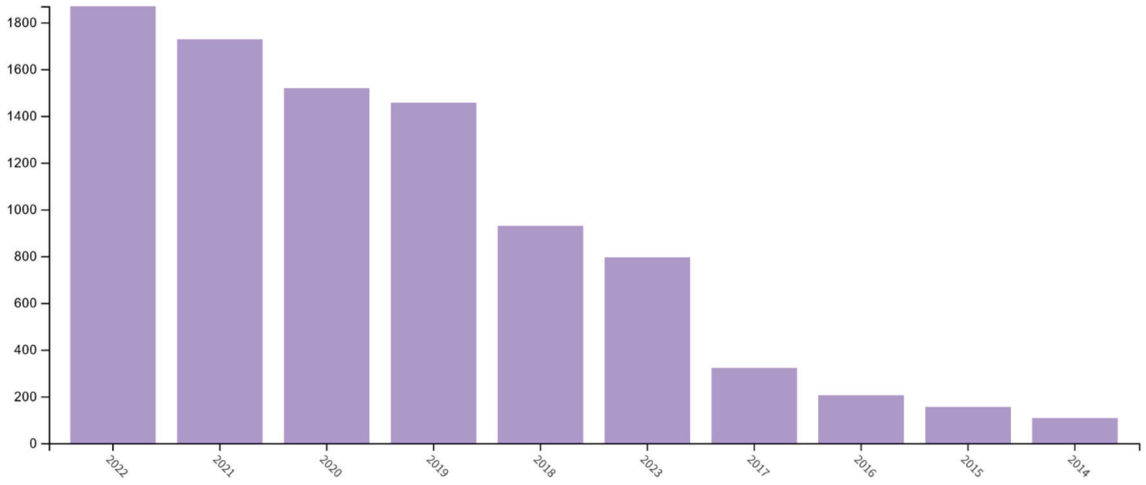


Figure 1. The frequency of publications from 2014 to 2022 in descending order. Source: Web of Science.

A total of 2080 of papers were published by the IEEE publisher, 1740 by Elsevier and 1272 by Springer Nature. A total of 4058 publications were included in the Computer Science main domain, 2908 in Business Economics, while 1571 in Engineering.

The other dataset used in this research consists of the monthly prices of Bitcoin extracted from Investing.com (accessed on 23 June 2023). The Bitcoin prices started to increase in 2017 and then again in 2019, but the highest jump was recorded in 2021. The price curve in 2021 had two humps followed by a steep valley in 2022 down to almost 15,000 USD. Today, prices are half of what they used to be in 2021 when they went up to almost 70,000 USD (as in Figure 2).



Figure 2. Analyzing the price evolutions of Bitcoin. Source: authors.

A rough estimation of the topics included in the abstracts can be obtained rapidly using WordCloud. The size of the words is directly proportional to their frequencies. The largest font size in Figure 3 belongs to blockchain, followed by Bitcoin, cryptocurrencies, transaction, network, protocol, system, application, etc. Using nltk and SpyCy Python libraries, pre-processing tasks were applied to the text. First, the punctuation was removed, letters were lowercased and regular stopwords plus several words specific to research papers were removed.

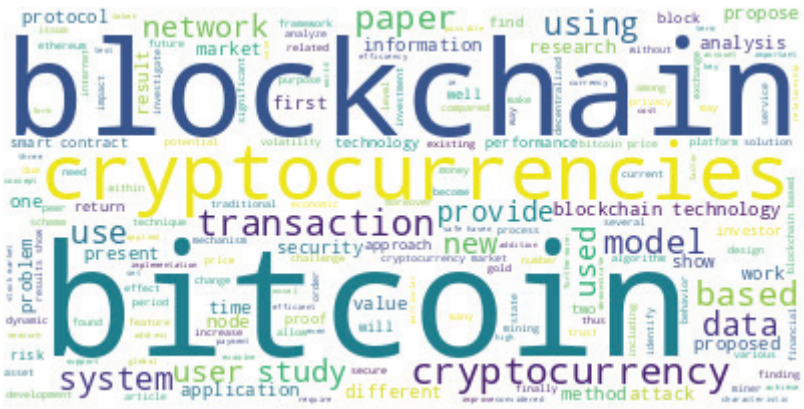


Figure 3. WordCloud applied to the abstracts of the research papers indexed in WoS. Source: authors.

The TextBlob library was used to assess the polarity of the abstracts. A percentage of 86.09% of the abstracts were positive, 12.73% were negative, while only 1.18% were neutral. However, the correlation between the abstract sentiment analysis and Bitcoin price (monthly records downloaded from Investing.com (https://www.investing.com/crypto/bitcoin/historical-data, accessed on 23 June 2023) was very weak (0.016). This weak correlation is maintained even if a lag of 12 months between the writing and publication of the research was considered.

To perform a prediction, the abstracts were vectorized. Both the BERT-transformer and word2vec were used in order to convert text into numeric vectors, but the prediction was far from the target. For prediction, five machine learning algorithms were implemented: the random forest, histogram gradient boosting, eXtreme Gradient Boosting, light gradient boosting and linear regression were combined using a voting regressor. However, the vectorized abstracts could not predict the Bitcoin prices. The prediction for the last 36 months is presented in Figure 4, showing a very low accuracy and the incapacity of the abstracts to play as an input variable for predicting Bitcoin prices. The same low accuracy level was obtained even when a 12-month or a 6-month lag between writing and publication was imposed.

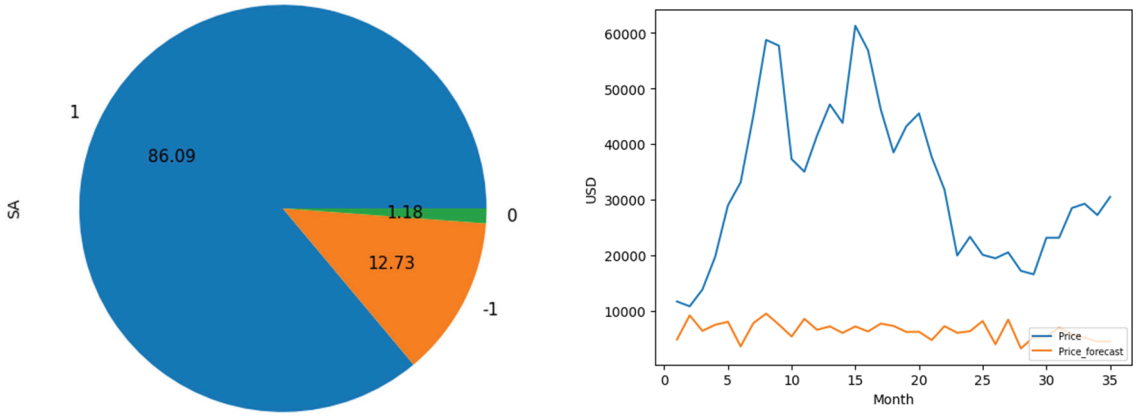


Figure 4. SA and prediction using vectorized abstracts with word2vec. Source: authors.

Initially, ten topics were set (K = 10), obtaining a topic coherence score = 0.303. Perplexity is another evaluation metric that signals how surprised a model is by new observations. It is computed as the normalized log-likelihood of a new observation, but recent studies have shown that perplexity is not sufficiently correlated with human judgment [77]. Thus, improving the perplexity may not lead to the human interpretation of the topics. This limitation shed light on another metric that is meant to better model human judgment: a topic coherence that combines a number of factors to assess the coherence between topics.

The baseline coherence score for the default LDA model is calculated, but furthermore, a sensitivity test is created to compute the optimal model hyperparameters: the number of topics (K), document-topic density—Dirichlet hyperparameter alpha: (α), and the word-topic density—Dirichlet hyperparameter beta: (β). Therefore, the optimal $\alpha = 0.01$, $\beta = 0.91$ with a corresponding $K = 6$ and $Cs = 0.385$, obtaining an improvement of 27.06%.

Using the genism and pyLDAvis libraries, six well-delimited topics with small overlaps are distinguished in the following LDA diagrams (Figure 5). The first and the largest topic that consisted in 32.1% of the tokens is undoubtedly related to *blockchain*, data, distributed/decentralized technology, system security, smart contracts and consensus (Figure 5a).

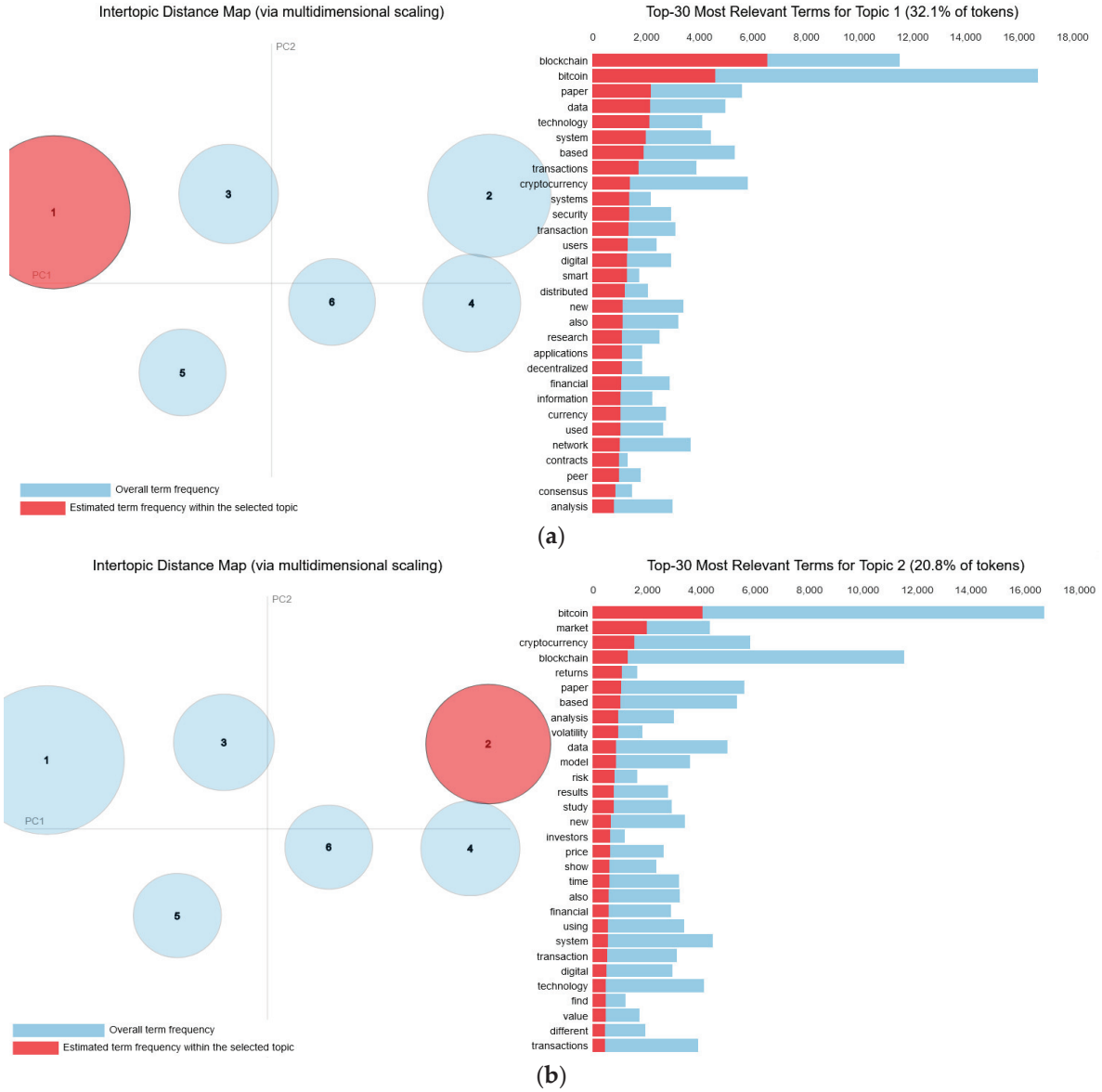
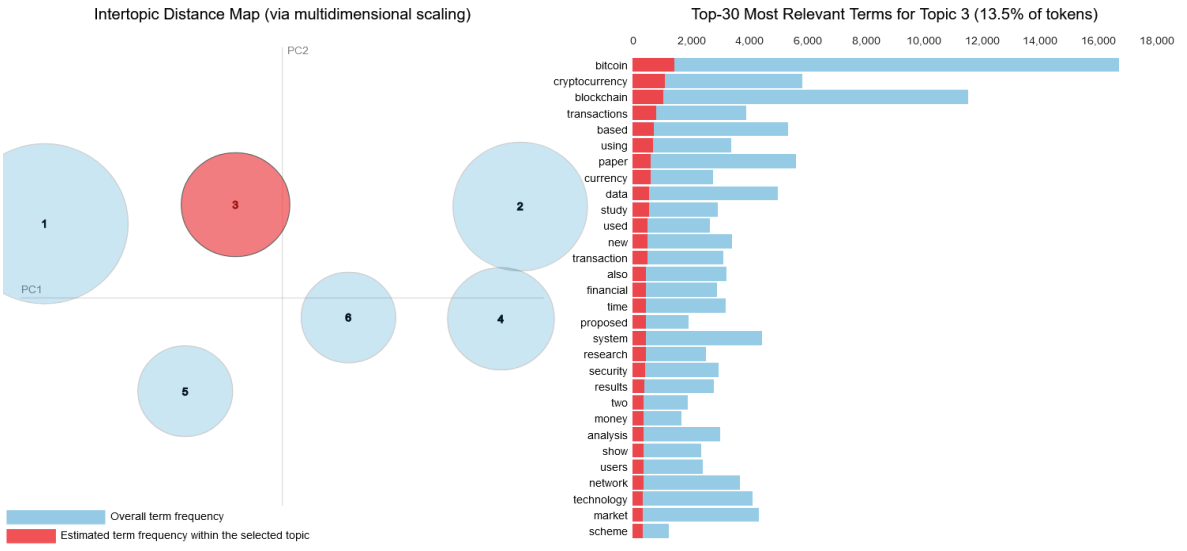
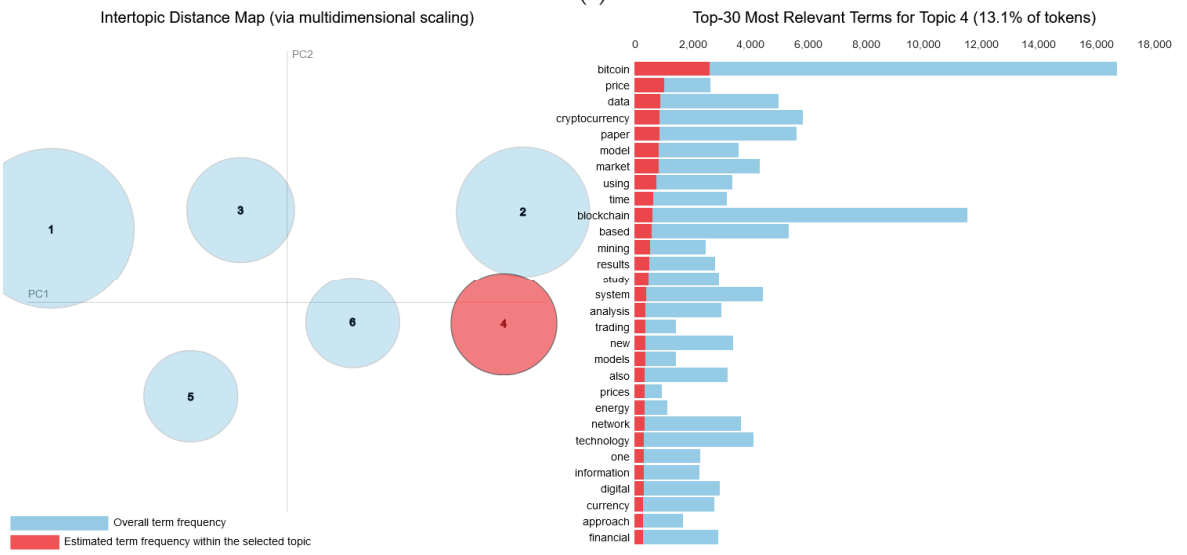


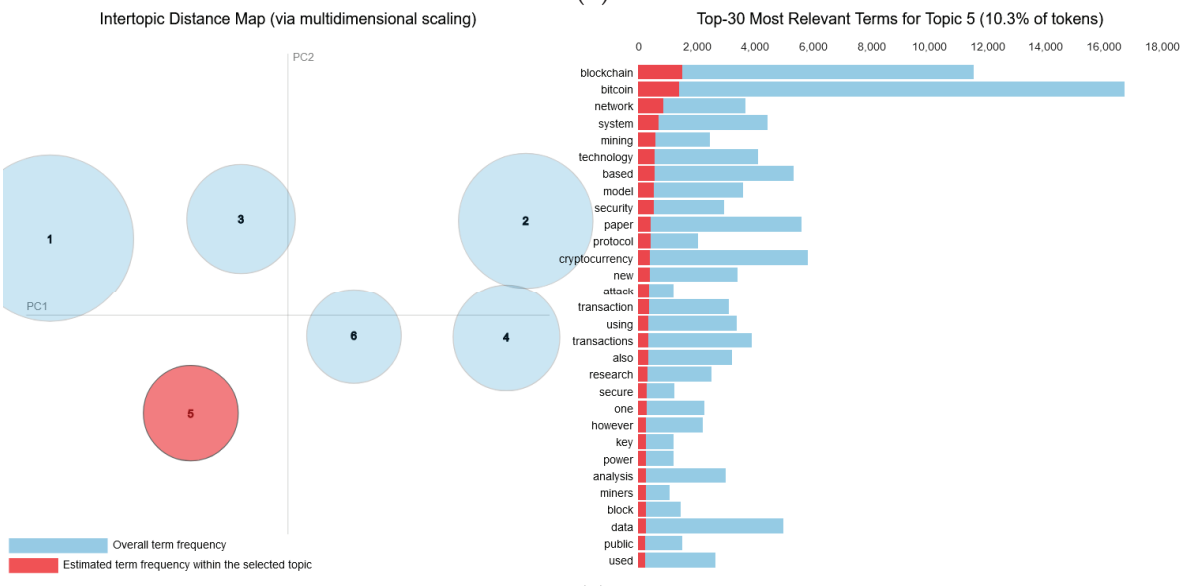
Figure 5. *Cont.*



(**c**)

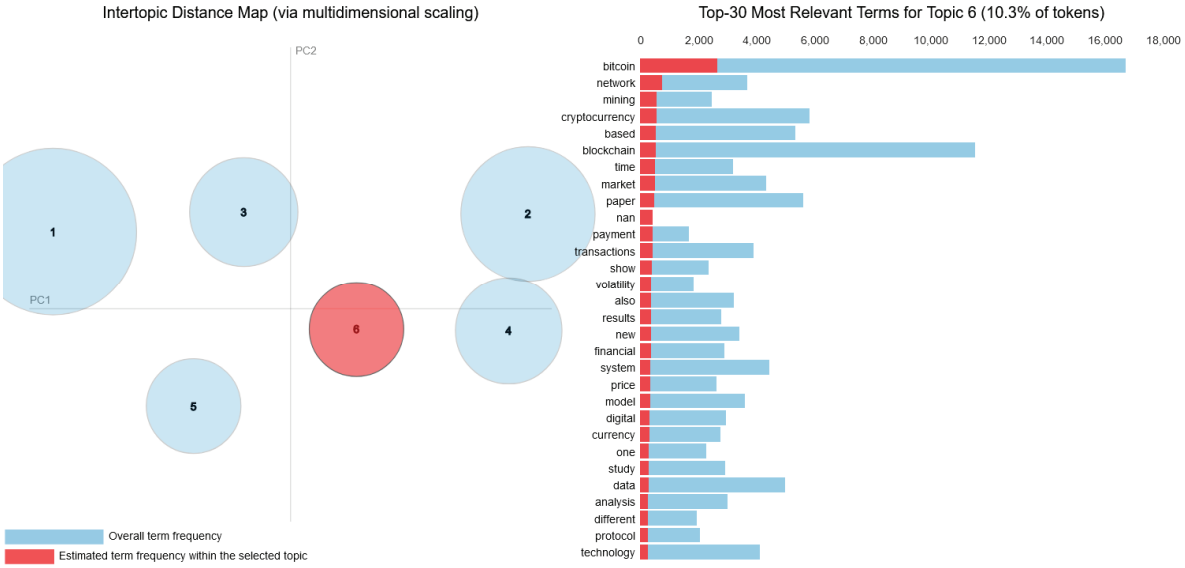


(**d**)



(**e**)

Figure 5. *Cont.*



(**f**)

Figure 5. LDA diagrams.

The second largest topic that consisted in 20.8% of the tokens focuses on the cryptocurrency *market*, returns, volatility, risk investors, financial aspects, price and value (Figure 5b). The third topic category that consisted in 13.5% of the tokens is related to cryptocurrency *transactions*, data, financial and system security (Figure 5c). The fourth topic that consisted in 13.1% of the tokens focuses on the Bitcoin *price*, pricing models, market, mining and energy (Figure 5d). The fifth topic that consisted in 10.3% is related to blockchain *network* security, mining technology, system security, attack, protocol, and miners (Figure 5e). The sixth topic is equal to the fifth topic in terms of tokens and is focused on the *mining* processing, market, payment, transactions, volatility, and financial issues (Figure 5f).

Topic modelling visualization is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. It is often used in text mining. In the following paragraphs, we break down what these visualizations generally represent: (1) *Inter-topic distance map* (via multidimensional scaling): The map shows different topics, represented as numbered circles. The size of each circle usually corresponds to the prevalence of the topic in the dataset. The distance between any two circles typically represents the similarity between the topics; closer circles suggest more similarity. The positioning along the axes indicates how the topics are spread; (2) *Top 30 most relevant terms for topics*: Each bar chart corresponds to a topic and lists the top 30 most relevant terms that characterize that topic. The length of the bars represents the term frequency within the topic. The red bars indicate terms that are more specific to the topic, while the blue bars indicate terms that are also relevant but not unique to the topic. The percentage of tokens indicates the weight of the topic in the overall dataset.

To compare, analyze, and interpret topic modelling visualizations, we look for patterns in the prevalence of topics, the uniqueness of terms to specific topics, and how topics might be related based on their proximity in the inter-topic distance map. The following aspects are analyzed to interpret subplots in Figure 5: (a) Comparing topic prevalence: The size of the circles in the distance map is checked to see which topics are more prevalent. Larger circles indicate a higher overall term frequency within the dataset; (b) Analyzing term relevance: To understand each topic, we review the terms listed in the bar charts. The terms further to the right are more relevant to the topic; (c) Identifying unique terms: The red bars represent terms that are not only frequent but also more specific to the topic, potentially giving a better sense of what distinguishes one topic from another; (d) Interpreting topic relationships: We examine which topics are close to each other on the map. Topics that are closer to each other might share some common terms or themes; (e) Topic contribution:

The percentage of tokens next to each bar chart title indicates how much of the text data was assigned to each topic, giving a sense of the topic's importance or dominance in the dataset. By following these steps, we construct a detailed understanding of the topics and how they relate to each other.

LDA is a form of unsupervised learning which is particularly well-suited for analyzing and categorizing large volumes of unlabeled text. It assumes that each document in a corpus can be described by a distribution of topics, and each topic can be described by a distribution of words. Each subplot in Figure 5 represents a visualization that come from LDA topic modelling. To analyze each subplot, we consider: (1) The inter-topic distance map: Each circle (bubble) represents a different topic discovered in the dataset. In the maps or subplots (a–f), the topics are numbered one through six. The size of a bubble reflects the prevalence of the topic. For instance, Topic 1 and Topic 2 are the largest bubble, suggesting they are prominent topics in the dataset. The distance between bubbles shows how different the topics are from each other; topics that are closer may share more common terms (Topic 2—market, and Topic 4—Bitcoin price), while those that are further apart are more distinct (Topic 2 and Topic 5—blockchain network security); (2) Top 30 most relevant terms: For each topic, there is a corresponding bar chart showing the 30 most relevant terms based on their frequency and distinctiveness. The x-axis indicates the frequency of terms, while the color coding (red to blue) represents how exclusive the terms are to the topic (with red being more exclusive). From the subplots in Figure 5, we further make several observations: (1) Topic size and spread: Larger topics (such as Topic 1 and Topic 2) have a wider spread of terms, indicating a broader discussion within the dataset. Smaller topics might represent more niche areas of discussion; (2) Term specificity: Red bars represent terms that are not only frequent but highly specific to the topic, giving us insight into what makes each topic unique; (3) Dominant terms: Common terms across different topics may represent overarching themes in the dataset, while unique terms give specific character to a topic; (4) Topic proximity: Topics that are close to each other on the distance map may have overlapping content or be related in some thematic way (only Topic 2 and Topic 4 indicate a small overlap, the rest of the topics being clearly distinct). As seen in Figure 5, in terms of the LDA results, the dataset contains a variety of discussions related to cryptocurrency, with some topics focusing on technical aspects, others on applications and on economic or trading perspectives.

5. Conclusions

More than 9100 research papers about Bitcoin and cryptocurrencies were published and indexed in Web of Science Clarivate platform until the end of July 2023. In this paper, we investigated the effect that publications may have on Bitcoin prices or the relationship between the academic community's opinion and Bitcoin prices. Furthermore, this paper aimed to uncover the main topics related to cryptocurrencies in general and Bitcoin in particular. LDA was applied to identify the relevant topics from the unstructured data of the research abstracts. The model hyperparameters were tuned in order to find the optimal number of topics.

Returning to the research questions exposed in the introduction, after merging and analyzing the two datasets, we are able to provide data-driven answers to (*RQ*1) What are the main topics investigated by the researchers in relation to cryptocurrencies and potential gaps? Blockchain, market, transactions, price, network security, and mining process are the six most predominant topics related to cryptocurrencies. Perplexity and topic coherence were calculated to optimally set the right number of topics.

To tune the number of topics and the parameters of the LDA method, the process proved to be time consuming (taking about 12 h) to process and test different scenarios. By tuning the LDA model, we obtained the optimal number of topics that provides the best coherence of topics and perplexity scores.

(RQ2) What is the academic community sentiment in relation to Bitcoin/cryptocurrencies? The abstracts are vectorized with BERT-transformers and word2vec and the potential of

abstracts to predict Bitcoin prices is analyzed. The sentiment analysis (SA) is performed and the relationship between SA and Bitcoin prices is showcased. Using the SA and investigating the polarity, we found that the majority of the abstracts conveyed a positive sentiment (86%), whereas 12% were negative and less than 2%. However, the findings indicate that academic publications have a minimal impact on Bitcoin prices, proving a weak link. This suggests that fluctuations in Bitcoin prices are not driven by scholarly research in this specific area. (RQ3) Is there a connection between academic publications and Bitcoin's price evolutions? Based on the Pearson correlation, the relation between the SA extracted from the abstracts and the Bitcoin prices is very weak or practically no evident connection could be detected. Further, the potential of the academic writing to influence the movements of Bitcoin prices is limited.

The six topics identified in this study suggest the mainstream or the relevant research on Bitcoin and cryptocurrencies. However, they also suggest gaps in the literature. For instance, the mining process that influences energy consumption and environmental aspects should be more investigated in the future. As Bitcoin accounts for roughly half of the energy demand of all cryptocurrencies, more energy-efficient consensus mechanisms are needed. Moreover, anti-money laundering measures, misbehaviors and crypto-crime related events should be further studied.

The practical implications of this study, which analyzed over 9100 research papers on Bitcoin and cryptocurrencies, are multifaceted:

(a) Guidance for future research. The identification of six predominant topics (blockchain, market dynamics, transactions, pricing, network security, and the mining process) highlights areas that have been the main focus of academic research. This can guide future studies to either build upon these areas or explore less-researched topics;
(b) Highlighting research gaps. The study points out gaps in the literature, especially regarding the environmental impact of the mining process and the need for more energy-efficient consensus mechanisms. This suggests a direction for new research endeavors, particularly in addressing the sustainability challenges of cryptocurrencies;
(c) Influence of academic research on the market. The finding that academic publications have minimal impact on Bitcoin prices provides valuable insights for investors, market analysts, and academics. It suggests that while academic research contributes to the understanding of cryptocurrencies, it does not significantly influence market movements;
(d) Sentiment analysis insights. The study reveals that the majority of academic sentiments toward Bitcoin/cryptocurrencies are positive. However, the weak correlation between this sentiment and Bitcoin prices can inform researchers and practitioners about the limited role of academic opinion in influencing market dynamics. Thus, while the study indicates that academic research does not significantly impact Bitcoin prices, it contributes significantly to the broader understanding of cryptocurrencies, highlighting key areas for future research and policy considerations.

One of the limitations of the current research is that it is confined to academic publications indexed in the Web of Science Clarivate platform. This may exclude relevant research published in other databases or platforms, potentially leading to a limited understanding of the entire body of cryptocurrency research. Another limitation is that the current research concentrates more on technical aspects like identifying topics, connections, and sentiment analysis, potentially underrepresenting economic interpretations and theories related to cryptocurrencies.

In future work, we aim to conduct our cryptocurrency study in conjunction with multiple social media and research platform data analyses, which may offer more timely insights, and intend to incorporate a deeper economic interpretation in sync with the dynamic pace of Bitcoin.

Author Contributions: Conceptualization, A.B. and S.-V.O.; methodology, A.B.; validation, A.B. and S.-V.O.; formal analysis, A.B. and S.-V.O.; investigation, A.B. and S.-V.O.; resources, A.B.; data

curation, A.B.; writing—original draft preparation, A.B. and S.-V.O.; writing—review and editing, A.B. and S.-V.O.; supervision, A.B.; project administration, S.-V.O. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was supported by Academy of Romanian Scientists, Ilfov 3, 050044 Bucharest, Romania, project title: "Soluții informatice pentru analiza impactului rețelelor de social media asupra instrumentelor investiționale cu grad ridicat de risc: cryptomonede și bursă".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data will be made available upon reasonable request.

Acknowledgments: This paper was supported by Academy of Romanian Scientists, Ilfov 3, 050044 Bucharest, Romania, project title: "Soluții informatice pentru analiza impactului rețelelor de social media asupra instrumentelor investiționale cu grad ridicat de risc: cryptomonede și bursă".

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Seow, K.T. Supervisory Control of Blockchain Networks. *IEEE Trans. Syst. Man Cybern. Syst.* **2019**, *50*, 159–171. [CrossRef]
2. Parino, F.; Beiró, M.G.; Gauvin, L. Analysis of the Bitcoin Blockchain: Socio-Economic Factors behind the Adoption. *EPJ Data Sci.* **2018**, *7*, 38. [CrossRef]
3. Liu, Y.; Zhang, L.; Zhao, Y. Deciphering Bitcoin Blockchain Data by Cohort Analysis. *Sci. Data* **2022**, *9*, 136. [CrossRef] [PubMed]
4. Prybila, C.; Schulte, S.; Hochreiner, C.; Weber, I. Runtime Verification for Business Processes Utilizing the Bitcoin Blockchain. *Futur. Gener. Comput. Syst.* **2017**, *107*, 816–831. [CrossRef]
5. Dias, R.; Alexandre, P.; Teixeira, N.; Chambino, M. Clean Energy Stocks: Resilient Safe Havens in the Volatility of Dirty Cryptocurrencies. *Energies* **2023**, *16*, 5232. [CrossRef]
6. Sapra, N.; Shaikh, I. Impact of Bitcoin Mining and Crypto Market Determinants on Bitcoin-Based Energy Consumption. *Manag. Financ.* **2023**, *49*, 1828–1846. [CrossRef]
7. Sharif, A.; Brahim, M.; Dogan, E.; Tzeremes, P. Analysis of the Spillover Effects between Green Economy, Clean and Dirty Cryptocurrencies. *Energy Econ.* **2023**, *120*, 106594. [CrossRef]
8. de Vries, A.; Gallersdörfer, U.; Klaaßen, L.; Stoll, C. Revisiting Bitcoin's Carbon Footprint. *Joule* **2022**, *6*, 498–502. [CrossRef]
9. Zaghloul, E.; Li, T.; Mutka, M.W.; Ren, J. Bitcoin and Blockchain: Security and Privacy. *IEEE Internet Things J.* **2020**, *7*, 10288–10313. [CrossRef]
10. BA, M. The Effect of Propagation Delay on the Dynamic Evolution of the Bitcoin Blockchain. *Digit. Commun. Netw.* **2019**, *6*, 157–166. [CrossRef]
11. Chawki, M. The Dark Web and the Future of Illicit Drug Markets. *J. Transp. Secur.* **2022**, *15*, 173–191. [CrossRef]
12. van Wegberg, R.; Oerlemans, J.-J.; van Deventer, O. Bitcoin Money Laundering: Mixed results? An Explorative Study on Money Laundering of Cybercrime Proceeds Using Bitcoin. *J. Financ. Crime* **2018**, *25*, 419–435. [CrossRef]
13. Böhmecke-Schwafert, M.; Moreno, E.G. Exploring Blockchain-Based Innovations for Economic and Sustainable Development in the Global South: A Mixed-Method Approach Based on Web Mining and Topic Modeling. *Technol. Forecast. Soc. Chang.* **2023**, *191*, 122446. [CrossRef]
14. Oprea, S.-V.; Bâra, A. Devising a Trading Mechanism with a Joint Price Adjustment for Local Electricity Markets Using Blockchain. Insights for Policy Makers. *Energy Policy* **2021**, *152*, 112237. [CrossRef]
15. Oprea, S.-V.; Bâra, A.; Andreescu, A.-I.; Cristescu, M.P. Conceptual Architecture of a Blockchain Solution for E-Voting in Elections at the University Level. *IEEE Access* **2023**, *11*, 18461–18474. [CrossRef]
16. Azzi, R.; Chamoun, R.K.; Sokhn, M. The Power of a Blockchain-Based Supply Chain. *Comput. Ind. Eng.* **2019**, *135*, 582–592. [CrossRef]
17. Grassman, R.; Bracamonte, V.; Davis, M.; Sato, M. Attitudes to Cryptocurrencies: A Comparative Study Between Sweden and Japan. *Rev. Socionetwork Strat.* **2021**, *15*, 169–194. [CrossRef]
18. Liu, F.; Wang, L.; Kong, D.; Shi, C.; Feng, Z.; Zhou, J.; Liu, J.; Li, Z. Is There More to Bitcoin Mining than Carbon Emissions? *Heliyon* **2023**, *9*, e15099. [CrossRef] [PubMed]
19. Kyriazis, N.A.; Papadamou, S.; Tzeremes, P. Are Benchmark Stock Indices, Precious Metals or Cryptocurrencies Efficient Hedges against Crises? *Econ. Model.* **2023**, *128*, 106502. [CrossRef]
20. Kyriazis, N.; Papadamou, S.; Tzeremes, P.; Corbet, S. The Differential Influence of Social Media Sentiment on Cryptocurrency Returns and Volatility during COVID-19. *Q. Rev. Econ. Financ.* **2022**, *89*, 307–317. [CrossRef]
21. Kyriazis, N.; Papadamou, S.; Tzeremes, P.; Corbet, S. Can Cryptocurrencies Provide a Viable Hedging Mechanism for Benchmark Index Investors? *Res. Int. Bus. Financ.* **2023**, *64*, 101832. [CrossRef]
22. Papadamou, S.; Kyriazis, N.A.; Tzeremes, P.G. Non-Linear Causal Linkages of EPU and Gold with Major Cryptocurrencies during Bull and Bear Markets. *North Am. J. Econ. Financ.* **2020**, *56*, 101343. [CrossRef]

23. Papadamou, S.; Kyriazis, N.A.; Tzeremes, P.; Corbet, S. Herding Behaviour and Price Convergence Clubs in Cryptocurrencies during Bull and Bear Markets. *J. Behav. Exp. Financ.* **2021**, *30*, 100469. [CrossRef]
24. Shahzad, M.K.; Bukhari, L.; Khan, T.M.; Islam, S.M.R.; Hossain, M.; Kwak, K.-S. BPTE: Bitcoin Price Prediction and Trend Examination Using Twitter Sentiment Analysis. In Proceedings of the 2021 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Republic of Korea, 20–22 October 2021; pp. 119–122.
25. Coulter, K.A. The Impact of News Media on Bitcoin Prices: Modelling Data Driven Discourses in the Crypto-Economy with Natural Language Processing. *R. Soc. Open Sci.* **2022**, *9*, 220276. [CrossRef]
26. Sharma, C.; Sharma, S.; Sakshi. Latent DIRICHLET Allocation (LDA) Based Information Modelling on BLOCKCHAIN Technology: A Review of Trends and Research Patterns Used in Integration. *Multimed. Tools Appl.* **2022**, *81*, 36805–36831. [CrossRef] [PubMed]
27. Liu, F.; Feng, Z.; Qi, J. A Blockchain-Based Digital Asset Platform with Multi-Party Certification. *Appl. Sci.* **2022**, *12*, 5342. [CrossRef]
28. Liu, F.; Fan, H.Y.; Qi, J.Y. Blockchain Technology, Cryptocurrency: Entropy-Based Perspective. *Entropy* **2022**, *24*, 557. [CrossRef]
29. Cotfas, L.-A.; Crăciun, L.; Delcea, C.; Florescu, M.S.; Kovacs, E.-R.; Molănescu, A.G.; Orzan, M. Unveiling Vaccine Hesitancy on Twitter: Analyzing Trends and Reasons during the Emergence of COVID-19 Delta and Omicron Variants. *Vaccines* **2023**, *11*, 1381. [CrossRef]
30. Kovacs, E.-R.; Cotfas, L.-A.; Delcea, C. January 6th on Twitter: Measuring Social Media Attitudes towards the Capitol Riot through Unhealthy Online Conversation and Sentiment Analysis. *J. Inf. Telecommun.* **2023**, 1–22. [CrossRef]
31. Yen, J.-C.; Wang, T. Stock Price Relevance of Voluntary Disclosures about Blockchain Technology and Cryptocurrencies. *Int. J. Account. Inf. Syst.* **2021**, *40*, 100499. [CrossRef]
32. Zhou, L.; Zhang, L.; Zhao, Y.; Zheng, R.; Song, K. A Scientometric Review of Blockchain Research. *Inf. Syst. E-Bus. Manag.* **2020**, *19*, 757–787. [CrossRef]
33. Alahi, I.; Islam, M.; Iqbal, A.; Bosu, A. Identifying the Challenges of the Blockchain Community from Stackexchange Topics and Trends. In Proceedings of the 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), Milwaukee, WI, USA, 15–19 July 2019.
34. Liu, F.; He, S.; Li, Z.; Xiang, P.; Qi, J.; Li, Z. An Overview of Blockchain Efficient Interaction Technologies. *Front. Blockchain* **2023**, *6*, 996070. [CrossRef]
35. Lade, M.; Welekar, R.; Dadiyala, C. Bitcoin Price Prediction and NFT Generator Based on Sentiment Analysis. *Int. J. Next-Gener. Comput.* **2023**, *14*. [CrossRef]
36. Zaman, S.; Yaqub, U.; Saleem, T. Analysis of Bitcoin's Price Spike in Context of Elon Musk's Twitter Activity. *Glob. Knowl. Mem. Commun.* **2022**, *72*, 341–355. [CrossRef]
37. Bashir, H.A.; Kumar, D. Investor Attention, Twitter Uncertainty and Cryptocurrency Market amid the COVID-19 Pandemic. *Manag. Financ.* **2022**, *49*, 620–642. [CrossRef]
38. Wang, F.; Gacesa, M. Semi-Strong Efficient Market of Bitcoin and Twitter: An Analysis of Semantic Vector Spaces of Extracted Keywords and Light Gradient Boosting Machine Models. *Int. Rev. Financ. Anal.* **2023**, *88*, 102692. [CrossRef]
39. Chursook, A.; Dawod, A.Y.; Chanaim, S.; Naktnasukanjn, N.; Chakpitak, N. Twitter Sentiment Analysis and Expert Ratings of Initial Coin Offering Fundraising: Evidence from Australia and Singapore Markets. *TEM J.* **2022**, *11*, 44–55. [CrossRef]
40. Gurrib, I.; Kamalov, F. Predicting Bitcoin Price Movements Using Sentiment Analysis: A Machine Learning Approach. *Stud. Econ. Financ.* **2021**, *39*, 347–364. [CrossRef]
41. Farimani, S.A.; Jahan, M.V.; Fard, A.M.; Tabbakh, S.R.K. Investigating the Informativeness of Technical Indicators and News Sentiment in Financial Market Price Prediction. *Knowl.-Based Syst.* **2022**, *247*, 108742. [CrossRef]
42. Hassan, M.K.; Hudaefi, F.A.; Caraka, R.E. Mining Netizen's Opinion on Cryptocurrency: Sentiment Analysis of Twitter Data. *Stud. Econ. Financ.* **2021**, *39*, 365–385. [CrossRef]
43. Dwivedi, D.N.; Vemareddy, A. Sentiment Analytics for Crypto Pre and Post Covid: Topic Modeling. In *International Conference on Distributed Computing and Intelligent Technology*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2023.
44. Burnie, A.; Yilmaz, E. An Analysis of the Change in Discussions on Social Media with Bitcoin Price. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019.
45. Ortu, M.; Vacca, S.; Destefanis, G.; Conversano, C. Cryptocurrency Ecosystems and Social Media Environments: An Empirical Analysis through Hawkes' Models and Natural Language Processing. *Mach. Learn. Appl.* **2021**, *7*, 100229. [CrossRef]
46. Ibba, G.; Vaccargiu, M. Analysis of Users' Most Discussed Topics and Trends on Blockchain Technologies and Smart Contracts. In Proceedings of the 2023 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER), Taipa, Macao, 21–24 March 2023.
47. Mulahuwaish, A.; Loucks, M.; Qolomany, B.; Al-Fuqaha, A. Topic Modeling Based on Two-Step Flow Theory: Application to Tweets about Bitcoin. *IT Prof.* **2023**, *25*, 52–63. [CrossRef]
48. Treiblmaier, H.; Swan, M.; de Filippi, P.; Lacity, M.; Hardjono, T.; Kim, H. What's Next in Blockchain Research? An Identification of Key Topics Using a Multidisciplinary Perspective. *ACM SIGMIS Database Database Adv. Inf. Syst.* **2021**, *52*, 27–52. [CrossRef]
49. Pathak, A.R.; Pandey, M.; Rautaray, S. Topic-Level Sentiment Analysis of Social Media Data Using Deep Learning. *Appl. Soft Comput.* **2021**, *108*, 107440. [CrossRef]

50. Schulte, M.; Eggert, M. Predicting Hourly Bitcoin Prices Based on Long Short-Term Memory Neural Networks. In *Innovation through Information Systems: Volume II: A Collection of Latest Research on Technology Issues*; Lecture Notes in Information Systems and Organisation; Springer: Cham, Switzerland, 2021.
51. Koutmos, D. Investor Sentiment and Bitcoin Prices. *Rev. Quant. Financ. Account.* **2022**, *60*, 1–29. [CrossRef]
52. Chen, J. Analysis of Bitcoin Price Prediction Using Machine Learning. *J. Risk Financ. Manag.* **2023**, *16*, 51. [CrossRef]
53. Bouri, E.; Salisu, A.A.; Gupta, R. The Predictive Power of Bitcoin Prices for the Realized Volatility of US Stock Sector Returns. *Financ. Innov.* **2023**, *9*, 62. [CrossRef] [PubMed]
54. Köse, N.; Ünal, E. The Asymmetric Effects of the Interest Rate on the Bitcoin Price. Financ. *Czech J. Econ. Financ.* **2023**, *73*, 189–217. [CrossRef]
55. Caliciotti, A.; Corazza, M.; Fasano, G. From Regression Models to Machine Learning Approaches for Long Term Bitcoin Price Forecast. *Ann. Oper. Res.* **2023**, 1–23. [CrossRef]
56. Liu, F.; Li, Z.; Jia, K.; Xiang, P.; Zhou, A.; Qi, J.; Li, Z. Bitcoin Address Clustering Based on Change Address Improvement. *IEEE Trans. Comput. Soc. Syst.* 2023; *early access*. [CrossRef]
57. Song, Y.; Chen, B.; Wang, X.-Y. Cryptocurrency Technology Revolution: Are Bitcoin Prices and Terrorist Attacks Related? *Financ. Innov.* **2023**, *9*, 29. [CrossRef] [PubMed]
58. Xinyu, W.; Shenghao, N.; Haibin, X. Forecasting Bitcoin Volatility Using Two-Component CARR Model. *Econ. Comput. Econ. Cybern. Stud. Res.* **2020**, *54*, 77–94. [CrossRef]
59. Hajek, P.; Hikkerova, L.; Sahut, J.-M. How Well Do Investor Sentiment and Ensemble Learning Predict Bitcoin Prices? *Res. Int. Bus. Financ.* **2023**, *64*, 101836. [CrossRef]
60. Jia, D.; Li, Y. Bounded Pool Mining and the Bounded Bitcoin Price. *Financ. Res. Lett.* **2023**, *52*, 103529. [CrossRef]
61. Bejan, C.A.; Bucerzan, D.; Crăciun, M.D. Bitcoin Price Evolution versus Energy Consumption; Trend Analysis. *Appl. Econ.* **2022**, *55*, 1497–1511. [CrossRef]
62. Aghashahi, M.; Bamdad, S. Analysis of Different Artificial Neural Networks for Bitcoin Price Prediction. *Int. J. Manag. Sci. Eng. Manag.* **2022**, *18*, 126–133. [CrossRef]
63. Jakubik, J.; Nazemi, A.; Geyer-Schulz, A.; Fabozzi, F.J. Incorporating Financial News for Forecasting Bitcoin Prices Based on Long Short-Term Memory Networks. *Quant. Financ.* **2022**, *23*, 335–349. [CrossRef]
64. Rathore, R.K.; Mishra, D.; Mehra, P.S.; Pal, O.; Hashim, A.S.; Shapi'I, A.; Ciano, T.; Shutaywi, M. Real-World Model for Bitcoin Price Prediction. *Inf. Process. Manag.* **2022**, *59*, 102968. [CrossRef]
65. Feng, W.; Zhang, Z. Currency Exchange Rate Predictability: The New Power of Bitcoin Prices. *J. Int. Money Financ.* **2023**, *132*, 102811. [CrossRef]
66. Cevik, E.I.; Gunay, S.; Dibooglu, S.; Yıldırım, D.Ç. The Impact of Expected and Unexpected Events on Bitcoin Price Development: Introduction of Futures Market and COVID-19. *Financ. Res. Lett.* **2023**, *54*, 103768. [CrossRef]
67. Ahmed, W.M.A. Robust Drivers of Bitcoin Price Movements: An Extreme Bounds Analysis. *North Am. J. Econ. Financ.* **2022**, *62*, 101728. [CrossRef]
68. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL HLT 2019—2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019.
69. Bello, A.; Ng, S.-C.; Leung, M.-F. A BERT Framework to Sentiment Analysis of Tweets. *Sensors* **2023**, *23*, 506. [CrossRef]
70. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the 1st International Conference on Learning Representations, ICLR 2013—Workshop Track Proceedings, Scottsdale, AZ, USA, 2–4 May 2013.
71. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. *Neural Inf. Process. Syst.* **2006**, *26*.
72. Pritchard, J.K.; Stephens, M.; Donnelly, P. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **2000**, *155*, 945–959. [CrossRef] [PubMed]
73. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022. [CrossRef]
74. Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X.; Jiang, X.; Li, Y.; Zhao, L. Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey. *Multimed. Tools Appl.* **2019**, *78*, 15169–15211. [CrossRef]
75. Bâra, A.; Oprea, S.-V. What Makes Electricity Consumers Change Their Behavior? Influence of Attitude and Perceived Impact of DR Programs on Awareness. *Kybernetes*, 2023; *ahead-of-print*. [CrossRef]
76. Oprea, S.-V.; Bâra, A. A Measurement Model for Electricity Consumers' Awareness with Covariance Structure Analyses. A Solid Pillar for Boosting Demand Response Programs. *Sustain. Energy Technol. Assess.* **2022**, *53*, 102738. [CrossRef]
77. Chang, J.; Boyd-Graber, J.; Gerrish, S.; Wang, C.; Blei, D.M. Reading Tea Leaves: How Humans Interpret Topic Models. In Proceedings of the Advances in Neural Information Processing Systems 22, Vancouver, BC, Canada, 7–10 December 2009.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Journal of
Theoretical and Applied
Electronic Commerce Research

MDPI

Article

Risk Analysis of Bankruptcy in the U.S. Healthcare Industries Based on Financial Ratios: A Machine Learning Analysis

Hadi Gholampoor [1] and Majid Asadi [2,*]

[1] Central Tehran Campus, Azad University, Tehran 1955847781, Iran
[2] College of Business, Northern Michigan University, Marquette, MI 49855, USA
* Correspondence: masadi@nmu.edu

Abstract: The prediction of bankruptcy risk poses a formidable challenge in the fields of economics and finance, particularly within the healthcare industry, where it carries significant economic implications. The burgeoning field of healthcare electronic commerce, continuously evolving through technological advancements and changing regulations, introduces additional layers of complexity. We collected financial data from 1265 U.S. healthcare industries to predict bankruptcy based on 40 financial ratios using multi-class classification machine learning models across various industry subsectors and market capitalizations. The exceptionally high post-tuning accuracy rates, exceeding 90%, along with high-performance metrics solidified the robustness and exceptional predictive capability of the gradient boosting model in bankruptcy prediction. The results also demonstrate the power and sensitivity of financial ratios in predicting bankruptcy based on financial ratios. The Altman models highlight the return on investment (ROI) as the most important parameter for predicting bankruptcy risk in healthcare industries. The Ohlson model identifies return on assets (ROA) as an important ratio specifically for predicting bankruptcy risk within industry subsectors. Furthermore, it underscores the significance of both ROA and the enterprise value to earnings before interest and taxes (EV/EBIT) ratios as important parameters for predicting bankruptcy based on market capitalization. Recognizing these ratios enables proactive decision making that enhances resilience. Our findings contribute to informed risk management strategies, allowing for better management of healthcare industries in crises like those experienced in 2022 and even on a global scale.

Keywords: bankruptcy; healthcare industry; USA stock market; financial ratio; machine learning; gradient boosting; Altman Z-score model; Ohlson model; financial health; risk management

1. Introduction

Bankruptcy risk analysis, a critical domain in economic and financial research, presents significant challenges due to the varied benchmarks across industries and their complex interactions with diverse factors in finance, law, and economics [1,2]. Particularly within the healthcare sector, these challenges are magnified due to their substantial influence on public health and socio-economic stability. This sector not only plays a pivotal role in employment and serves as a key focus for investment but also substantially impacts the broader medical and financial ecosystems through the repercussions of hospital and health insurance provider bankruptcies [1–10]. Moreover, the correlation between healthcare sector financial distress and personal bankruptcies, notably driven by uninsured medical expenses, underscores this issue's severity; nearly two-thirds of all U.S. personal bankruptcies are medically related [7,8]. Additionally, the interaction between instabilities in the healthcare industry bankruptcies has been recognized [11]. In this context, the burgeoning field of electronic commerce (e-commerce) emerges as a significant factor. The evolution and integration of e-commerce have revolutionized investment patterns and financial stability across many sectors, including healthcare. E-commerce platforms enhance the accessibility and distribution of healthcare services and products, thereby influencing

the financial dynamics of healthcare institutions. These platforms also affect consumer behavior and spending on health services, potentially increasing the financial risk for healthcare providers that fail to adapt to digital marketplaces. Furthermore, e-commerce introduces new regulatory challenges and competitive pressures that can exacerbate financial instability in the healthcare sector. As e-commerce continues to expand, its impact on the stock market dynamics that underpin the financial health of medical entities cannot be overlooked. Therefore, the interplay between e-commerce and healthcare financial stability is critical for predicting potential bankruptcies in this sector. This study aims to address this nuanced landscape by focusing on the bankruptcy risk in U.S. medical and healthcare entities, influenced by stock market dynamics.

The year 2022 presents a unique context for analyzing bankruptcy risk in the U.S. healthcare industries, characterized by challenging economic conditions. This period saw a bear market, record-high inflation levels not seen in over four decades, and a marked increase in interest rates [Federal Reserve, 2023], all unfolding alongside the ongoing recovery from the COVID-19 pandemic. Such conditions precipitated significant market volatility, acutely impacting growth-sensitive sectors, notably medical and healthcare. Our research is positioned to delve into bankruptcy risk within this particular timeframe, aiming to provide healthcare stakeholders with critical insights. By analyzing this period, we intend to alert them to potential financial instabilities and propose preemptive strategies to effectively navigate and mitigate bankruptcy risks during comparable periods of economic adversity.

2. Literature Review

Given the capability of financial ratios to reflect critical elements of a company's financial status, several studies have investigated the use of these ratios in assessing bankruptcy risk, particularly in sectors outside of medical and healthcare. These studies typically employ traditional statistical methods, such as discriminant and regression analysis, to analyze bankruptcy risks. Supriyanto et al. conducted an investigation into financial distress in mining companies, with a specific focus on financial ratios [12]. Similarly, Amalia et al. examined financial ratios to predict company bankruptcy in the cigarette industry [13]. Lee et al. explored the financial ratios of savings banks in relation to bankruptcy through quantitative empirical analysis using statistical models [14]. Additionally, Tian et al. studied bankruptcy prediction across international markets with financial ratios by employing adaptive statistical techniques and a discrete hazard statistical model [15]. Traditional statistical methods can generally identify the significance and relationship of each individual parameter, such as financial ratios, with bankruptcy. Nevertheless, given the multi-factorial nature of bankruptcy, it is imperative to consider the impact of all financial ratios simultaneously for bankruptcy prediction. Moreover, the intricate interconnections between these ratios should also be taken into account during this predictive analysis. Therefore, these traditional statistical analyses may not be fully capable of discerning the intricate interplay between financial parameters and bankruptcy [16].

Artificial intelligence and machine learning methods have demonstrated a greater aptitude than traditional statistical analysis for addressing concerns in evaluating bankruptcy risk. Having emerged as pivotal technologies in the fourth industrial revolution [17,18], these methods enable the development of sophisticated models that consider both the multifaceted roles of financial ratios and the interrelationships between them. Recent studies have delved into the application of diverse machine learning algorithms to ascertain which models are most effective in predicting bankruptcy and financial distress [19–22]. These investigations aim to enhance the accuracy and reliability of bankruptcy predictions by leveraging the advanced analytical capabilities of machine learning. Additionally, some of these studies have applied these models to assess the dynamics of financial distress and bankruptcy in non-healthcare industries across various subsectors [23–35]. Carmona et al. implemented innovative machine learning algorithms to predict bankruptcy in French firms [36]. They assessed the prediction of business financial distress and bankruptcy in French firms using a novel machine learning model. In another study, they proposed

an enhanced gradient boosting machine learning algorithm for bankruptcy prediction, and their experimental results substantiated its superiority in comparison to traditional feature selection methods [37]. Lombardo et al. contributed to the prediction of corporate bankruptcy by utilizing machine learning algorithms that estimate survival probabilities and predict defaults, employing time-series accounting data [38]. They also suggested an innovative method capable of detecting emerging risks within particular sectors and industries, thus aiding in the identification of market segments where firms remain undisturbed by disruptions [39]. Bragoli et al. also developed a predictive model that classified solvent and bankrupt firms, utilizing the enhanced gradient boosting machine learning algorithm [40].

Some studies have concentrated on predicting bankruptcy risk during specific critical time periods characterized by economic crises. This focus is crucial as economic downturns significantly alter the financial landscape, thereby impacting the accuracy and relevance of bankruptcy prediction models. Liu et al. focused on high-tech and startup businesses in Europe after the 2008 economic crisis, emphasizing the predictive capabilities of machine learning, which could potentially even contribute to preventing failures or acquisitions [41]. Papík et al. explored the prediction of financial distress and bankruptcy risk among small- and medium-sized enterprises (SMEs) during the COVID-19 crisis using machine learning [42]. They advised against the use of qualitative indicators in SME prediction models and noted a shift in the relationship with historical financial data between 2019 and 2020 due to the COVID-19 pandemic. Liu et al. explored the influence of public health crises on the economic prospects of SMEs by utilizing machine learning analysis [43]. Their investigation into the predictive aspects of credit default risks for SMEs post COVID-19 underscores the effectiveness of machine learning methods in assessing and controlling credit risk, thereby elucidating its impact on SMEs.

3. Objective

Previous research has laid important groundwork in bankruptcy risk analysis; however, unresolved issues persist, and new challenges have emerged, particularly in the context of the economic climate of 2022 [1,38]. There is considerable uncertainty regarding the most effective financial metrics for predicting bankruptcy in the healthcare industry. Notably, the application of bankruptcy prediction within the healthcare industry, especially through machine learning techniques, remains underexplored. This study innovatively focuses on predicting bankruptcy across various subsectors and market capitalizations within the healthcare industry. We utilized a comprehensive set of financial ratios for prediction, acknowledging their widespread applicability and establishing a new precedent in financial analysis. Financial ratios offer a diverse set of financial metrics that provide a concise yet powerful snapshot of a company's financial health and resilience [44]. This research aims to address these gaps by focusing on the bankruptcy risk analysis of healthcare industries in the United States, with particular emphasis on the economically tumultuous year of 2022. Our primary objective was to employ machine learning algorithms to analyze the intricate relationship between financial ratios and bankruptcy risk in the healthcare industry. Additionally, we aimed to identify the most effective financial ratios that are most crucial and sensitive for predicting bankruptcy in these sectors. Our approach seeks to provide strategic insights for stakeholders in the healthcare stock market, including investors, policymakers, shareholders, and healthcare professionals. Through this, we aspire to contribute to the financial stability and resilience of the medical and healthcare sectors by facilitating informed decision making.

4. Data, Methodology, and Variables

4.1. Data Collection

Due to the lack of precise financial information on the medical and healthcare stock market in available financial databases, we gathered financial data from the U.S. healthcare stock market via Sec.gov, covering 1265 companies. The data were sourced from official 2022

reports, including balance sheets, income statements, and cash flow statements, all publicly available. For the purpose of this study, the time-dependent financial ratios were derived from data spanning a full calendar year, specifically from 1 January 2022 to 1 January 2023. This approach was deliberately chosen to align with standard financial reporting practices, which typically reflect the financial status of an entity at both the beginning and the end of each year. The calculated ratios in Table 1 encapsulate the financial dynamics within this period, offering a snapshot of each company's financial health at these critical fiscal junctures. The detailed formulas and definitions for these 40 financial ratios are provided in Table 1 and Supplementary Material S1. These ratios encompass a multitude of nuanced sub-parameters across key financial dimensions, including but not limited to indebtedness, cash flow dynamics, profitability metrics, and sector-specific peculiarities, all of which are fundamentally extrapolated from an extensive array of financial data pertaining to stocks. For instance, the debt-to-equity ratio sheds light on a company's leverage structure, whereas ratios such as the operating cash flow ratio, price-to-free cash flow ratio, and the cash ratio collectively provide a multifaceted understanding of its liquidity situation. In a similar vein, the current ratio and the debt ratio deliver a concise yet informative overview of the company's debt landscape. On the front of profitability analysis, indicators such as the gross profit margin, net profit margin, and return on equity are of paramount importance. Further, ratios like price-to-sales and price-to-book value are essential in gleaning insights on the peculiarities inherent to the specific industry in question. Collectively, these ratios are imperative in constructing a comprehensive and insightful portrayal of a firm's financial vitality and its strategic standing within its respective sector.

Table 1. The formulas for 40 financial ratios (features) used to evaluate the bankruptcy risk of healthcare industries in the U.S. The symbol # represents the number of financial ratios (FR number).

#	Ratio Name	Formula	#	Ratio Name	Formula
1	Earnings ratio	$\frac{Net\ income - Preferred\ dividends}{Common\ outstanding\ shares}$	21	Price-to-operating-cash-flow ratio	(Market capitalization)/(Cash flow from operation)
2	Price-to-earnings ratio	(Market price per share)/(Earning per share)	22	Book value per share	$\frac{Shareholders'\ equity - Preferred\ equity}{Outstanding\ shares}$
3	Price-to-sale ratio	(Market cap)/(Total sale)	23	Liquidity ratio	(Current assets)/(Current liabilities)
4	Price-to-book ratio	(Market price per share)/(Book value per share)	24	Debt ratio	(Total debt)/(Total assets)
5	Earnings per share	(Net income)/(Number of outstanding shares)	25	Price-to-cash-flow ratio	(Market price per share)/(Cash flow per share)
6	Return on equity	(Net income)/(Shareholders' equity)	26	Cash ratio	(Cash and cash equivalents)/(Current liabilities)
7	Return on assets	(Net income)/(Total assets)	27	Return on capital employed	(Operating profit)/(Capital employed)
8	Return on investment	(Net income)/(Cost of investment)	28	Return on sales	(Operating profit)/(Net sales)
9	Gross margin	(Gross profit)/Revenue	29	Cash return on invested capital	(Free cash flow)/(Invested capital)
10	Net margin	(Net profit)/Revenue	30	Cash return on equity	(Free cash flow)/(Shareholders' equity)
11	Operating margin	(Operating income)/Revenue	31	Cash return on assets	(Free cash flow)/(Total assets)
12	Current ratio	(Current assets)/(Current liabilities)	32	Free cash flow margin	(Free cash flow)/(Total revenue)

Table 1. *Cont.*

#	Ratio Name	Formula	#	Ratio Name	Formula
13	Quick ratio	$\frac{Current\ assets - inventory}{Current\ liabilities}$	33	Operating cash flow margin	(Operating cash flow)/(Total revenue)
14	Debt-to-equity ratio	Debt/(Shareholder's equity)	34	Total asset turnover ratio	Revenue/(Average total assets)
15	Accounts receivable turnover ratio	(Net credit sales)/(Average account receivable)	35	Equity ratio	(Total equity)/(Total assets)
16	Operating cash flow ratio	(Cash from operations)/(Current liabilities)	36	Excess return	(Actual return)—(Required return)
17	EV-to-EBITDA ratio	(Enterprise value)/EBITDA	37	Fixed asset turnover ratio	(Net sales)/(Average fixed assets)
18	EV-to-sales ratio	(Enterprise value)/Sales	38	Receivables turnover ratio	(Net sales)/(Average account receivable)
19	EV-to-EBIT ratio	(Enterprise value)/EBIT	39	Total-liabilities-to-total-assets ratio	(Total liabilities)/(Total assets)
20	Price-to-free-cash-flow ratio	(Market capitalization)/(Free cash flow)	40	EBIT margin	EBIT/Revenue

4.2. Study Design and Machine Learning Analysis

This study employed a machine learning classification approach to analyze bankruptcy risk. Forty financial ratios (detailed in Table 1) served as features within the machine learning methodology, with the bankruptcy value for each of the 1265 industries as the prediction target. To ensure rigorous computation of these targets (bankruptcies), we utilized three widely recognized empirical bankruptcy prediction models: the Altman Z-score model, the modified Altman model, and the Ohlson model—commonly referred to as the O-score. These models have been empirically validated in numerous studies for their effectiveness in predicting bankruptcy risk, providing a robust and reliable framework for the analysis [45–49]. The formulations for these three bankruptcy models are provided in Equations (1)–(3) [50–53].

$$\text{Altman Z-score} = 1.2P1 + 1.4P2 + 3.3P3 + 0.6P4 + 1.0P5 \quad (1)$$

P1: Working capital/Total assets;
P2: Retained earnings/Total assets;
P3: Earnings before interest and taxes (EBIT)/Total assets;
P4: Market value of equity/Total liabilities;
P5: Sales/Total assets.

$$\text{ModifiedAltman} = 6.56K1 + 3.26K2 + 6.72K3 + 1.05K4 \quad (2)$$

K1: Working capital/Total assets;
K2: Retained earnings/Total assets;
K3: EBIT/Total assets;
K4: Book value of equity/Total liabilities.

$$\text{Ohlson} = -1.32 - 0.407\,X1 + 6.03X2 - 1.43X3 + 0.0757X4 - 2.37X5 - 1.83X6 + 0.285X7 - 1.72X8 - 0.521X9 \quad (3)$$

X1: Log (total assets/(gross national product—level price index));
X2: Total liabilities/total assets;
X3: Working capital/total assets;
X4: Current liabilities/current assets;

X5: Net income/total assets;

X6: Cash flow from operations/total liabilities;

X7: 1 if net negative income; 0 if otherwise;

X8: 1 if the total liabilities > total assets; 0 if otherwise;

X9: (Net income T—Net income T-1)/(Net income T + Net income T-1); "T" refers to a specific time period.

The machine learning classification algorithms used to predict bankruptcy were logistic regression, decision tree, naive Bayes, K-nearest neighbors (KNN), and support vector machine (SVM), along with ensemble learning algorithms including adaptive boosting (AdaBoost), gradient boosting, and random forest algorithms. A comprehensive explanation of the machine learning algorithms employed in the current study, along with their corresponding equations, is provided in Supplementary Material S2.

Market capitalization serves as a powerful indicator of a company's size, reflecting not only the equity value perceived by the market but also providing insight into the financial robustness of the firms. Previous studies have shown that the market capitalization of industries is of particular importance in the analysis of bankruptcy risks [54–56]. On the other hand, in examining bankruptcy risk in the healthcare industries, industry subsector classification plays a pivotal role, as each subsector contributes distinct elements to the overall risk profile. Given the significance of both market capitalization and industry subsectors in bankruptcy risk analysis, this study employs two approaches: Bankruptcy risks are analyzed based on the industry subsector and the market capitalization of industries. This dual perspective enables a nuanced assessment of how industry dynamics and company size characteristics interact to influence bankruptcy likelihood. To implement these approaches, we applied two distinct multi-class machine learning classifications. Firstly, the targets were classified into four different industry subsector classes. Secondly, the targets were classified into five different market capitalization classes. It is important to note that we had three targets (Altman Z-Score, modified Altman, and Ohlson bankruptcy models), and as a result, we repeated the entire machine learning multi-class classification procedure three times. Finally, the algorithm's predictions were compared to the test outcomes using confusion matrices to evaluate accuracy, precision, recall, and F1 scores. The algorithm with the best performance metrics was selected as the final method for bankruptcy prediction. It is worth mentioning that all targets were not completely balanced. Hence, we employed stratified sampling in the train–test split to preserve the target variable bankruptcy distribution in both training and testing sets. Additionally, StratifiedKFold cross-validation was used, demonstrating the suitability of our chosen methodology for maintaining the proportion of classes across folds. These primary techniques effectively solved issues related to minor imbalances during the prediction process, making advanced methods such as over-sampling or under-sampling unnecessary. The machine learning code was implemented using Python software, version 3.11.

It should be noted that the four industry subsector classes, defined based on the healthcare sectors of the U.S. industries, include (1) biotechnology; (2) medical, which comprises medical devices, instruments and supplies, and medical distribution; (3) pharmacology, encompassing drug manufacturers—specialty and generic, as well as general—and pharmaceutical retailers; and (4) healthcare, which covers diagnostics and research, health information services, medical care facilities, and healthcare plans. The five market capitalization classes were defined as follows [57]: (1) large-cap, with a market value of USD 10 billion or more; (2) mid-cap, with a market value between USD 2 billion and USD 10 billion; (3) small-cap, with a market value between USD 300 million and USD 2 billion; (4) micro-cap, with a market value between USD 50 million and USD 300 million; and (5) nano-cap, with a market value of less than USD 50 million.

This study recognizes that dividing industries into various subsectors and market capitalizations has enabled a multi-class classification that includes some minority classes. Recent research suggests that for non-medical data, such as financial datasets, resampling methods can effectively address the prediction challenges associated with minority

classes [58]. However, the research also advises that, in most cases, standard classifiers are preferable for non-medical datasets similar to those analyzed in this study. Based on these insights, this study employs standard classifiers and does not utilize resampling or cost-sensitive methods.

4.3. Statistical Analysis

For the assessment of baseline characteristics within our dataset, we employed statistical analysis methods and computed measures, including the mean and standard error (SE), for 40 distinct financial ratios. The Shapiro–Wilk test was applied to determine the normality of the distribution within our dataset. To ascertain the statistical significance of individual features, we utilized the one-way analysis of variance (ANOVA). A p-value threshold of 0.05 was set for determining statistical significance.

5. Results

Table 2 presents a summary of the financial ratios, including the mean, SE, and p-values, which underscore the observed differences in the financial ratios. A three-dimensional principal component analysis (PCA) was conducted on the dataset, comprising industries and their corresponding financial ratios (Figure 1). The resulting graphical panels highlight variations in the distribution of industries, categorized by industry subsectors and market capitalization tiers. This method of dimensionality reduction effectively preserves the inherent variance within the dataset, facilitating a more nuanced examination of the complex attributes of the industries [59]. Within the derived three-dimensional space, individual industries are represented as points, with their positions reflecting the synthesized financial ratios. The results, as depicted in Figure 2, reveal a significant bankruptcy risk for healthcare industries in the U.S. during 2022. The bankruptcy risk, calculated based on the Altman Z-Score, modified Altman, and Ohlson models, is depicted at 88.5%, 62.1%, and 43.2%, respectively, as shown in the non-green segments of Figure 2. The observed disparities in bankruptcy risk evaluations are attributable to the distinct methodological underpinnings of the Altman models and Ohlson model. The Altman models, originally calibrated for manufacturing firms, predominantly rely on financial ratios reflective of this sector's characteristics. In contrast, the Ohlson model boasts a design intended for broad applicability across diverse corporate environments, such as service-centric industries, and integrates an extensive array of financial ratios.

Table 2. Statistical summary of financial ratios (features) including mean, SE, and p-value. FR, financial ratio; SE, standard error.

Financial Ratios	Mean	SE	p-Value	Financial Ratios	Mean	SE	p-Value
FR 1	−7.35	1.93	0.0001	FR 21	−600.31	1417.46	0.6720
FR 2	1.66	11.78	0.8876	FR 22	7.46	0.75	0.0000
FR 3	1241.45	587.43	0.0347	FR 23	6.44	0.23	0.0000
FR 4	330,027.84	270,878.66	0.2233	FR 24	0.18	0.01	0.0000
FR 5	−7.35	1.93	0.0001	FR 25	−9.50	8.61	0.2701
FR 6	−1.01	1.22	0.4117	FR 26	2.64	0.09	0.0000
FR 7	−0.55	0.03	0.0000	FR 27	−0.72	0.20	0.0002
FR 8	9.15	2.41	0.0001	FR 28	−32.28	3.94	0.0000
FR 9	0.50	0.03	0.0000	FR 29	1.75	1.18	0.1396
FR 10	−32.43	4.01	0.0000	FR 30	1.63	1.62	0.3141
FR 11	−32.28	3.94	0.0000	FR 31	−0.40	0.02	0.0000
FR 12	6.44	0.23	0.0000	FR 32	−27.71	3.52	0.0000
FR 13	3.94	0.07	0.0000	FR 33	−27.28	3.39	0.0000
FR 14	−0.48	0.43	0.2664	FR 34	0.38	0.02	0.0000
FR 15	−46.13	18.83	0.0144	FR 35	0.51	0.01	0.0000
FR 16	−2.35	0.09	0.0000	FR 36	−0.43	0.01	0.0000
FR 17	8.40	6.88	0.0000	FR 37	1.59	0.13	0.0000
FR 18	1651.01	586.73	0.0049	FR 38	−46.13	18.83	0.0144
FR 19	4.92	3.10	0.1124	FR 39	0.49	0.01	0.0000
FR 20	−0.34	0.17	0.0495	FR 40	−32.30	3.94	0.0000

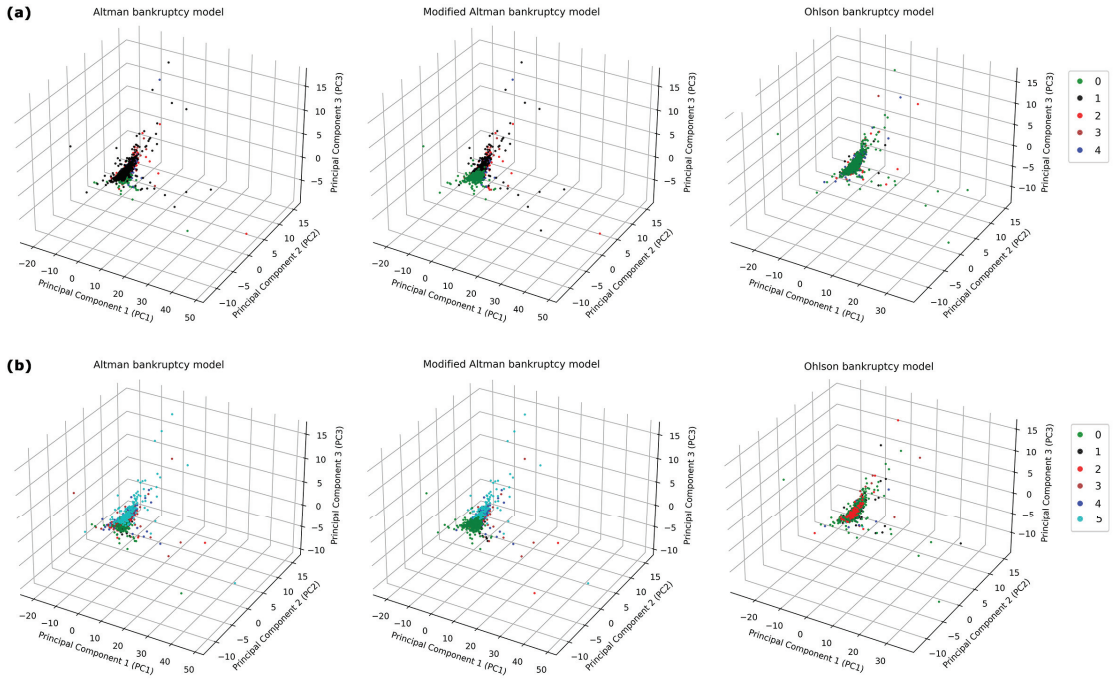


Figure 1. Three-dimensional principal component analysis (PCA) of U.S. healthcare industries features: This figure illustrates the PCA plots of 1265 U.S. healthcare industries, visualized in three-dimensional space through the first three principal components. This analytical approach is used to identify patterns regarding (**a**) bankruptcy across different medical industry subsectors and (**b**) bankruptcy relative to market capitalization categories. Each data point represents a unique stock, with its spatial positioning derived from an aggregation of features within the dimensions defined by principal components 1, 2, and 3. In panels (**a**), industry subsector classes are labeled as 0 for solvency and from 1 to 4 for bankruptcy in biotechnology, medical devices, pharmacology, and healthcare services, respectively. In panels (**b**), market capitalization classes are defined as 0 for solvency and from 1 to 5 for bankruptcy in large-cap, mid-cap, small-cap, micro-cap, and nano-cap categories, respectively.

In our investigation, StratifiedKFold cross-validation was systematically applied to each classifier, ensuring that the folds preserved a proportional distribution of classes. This procedure, in conjunction with detailed hyperparameter optimization—as documented in Supplementary Material S3—was critical in reducing the likelihood of data leakage. The application of these meticulous stratification and tuning strategies led to enhancements in models' predictive accuracy, with increases ranging from 7.1% to 10.3%. In addition, all models exhibited acceptable accuracy post tuning, with the sole exception of the naive Bayes classifier (Supplementary Material S3). It should be noted that the naive Bayes model, despite its probabilistic foundation and advantages in simplicity and computational speed, did not achieve as high an accuracy as the other models in predicting bankruptcy within the U.S. healthcare industries. The relatively lower performance of this model may be explained by the model's assumption of independence among predictors, a condition presumably not met by the complex interactions inherent in our dataset. The consistency of performance improvements post tuning across the remaining models substantiates the robustness of our analytical framework and corroborates the nonexistence of overfitting within our results. Our final tuned models exhibited acceptable accuracy rates for predicting bankruptcy using financial ratios inspired by the Altman Z-score model, with performance by classifier as follows: SVM (87.3%), KNN (82.3%), naive Bayes (68.0%), logistic regression (86.3%), decision tree (86.4%), AdaBoost (87.3%), random forest (90.6%), and gradient boosting (90.8%), as detailed in Supplementary Material S3 for industry subsectors. The accuracies for the modified Altman bankruptcy model were 88.1%, 83.3%, 66.7%, 87.3%,

87.4%, 84.1%, 90.3%, and 90.6%, respectively, while the Ohlson model showed comparable performance with accuracies of 87.9%, 80.0%, 69.1%, 87.8%, 88.2%, 74.5%, 89.3%, and 90.0%. Classifications based on market capitalization also yielded acceptable accuracy rates, broadly aligning with those based on industry subsectors (Supplementary Material S3). The results depicted indicate that, in addition to accuracy, the precision, recall, and F1 scores for the models fall within a satisfactory range. Finally, the results of our analysis demonstrated that both the random forest and gradient boosting algorithms exhibited exceptional accuracy and robustness in predicting bankruptcy based on all three bankruptcy models utilized. Upon closer examination of the results, the gradient boosting model was found to have a marginally higher accuracy compared to the random forest algorithm. Therefore, we chose to focus our analysis on this particular algorithm.

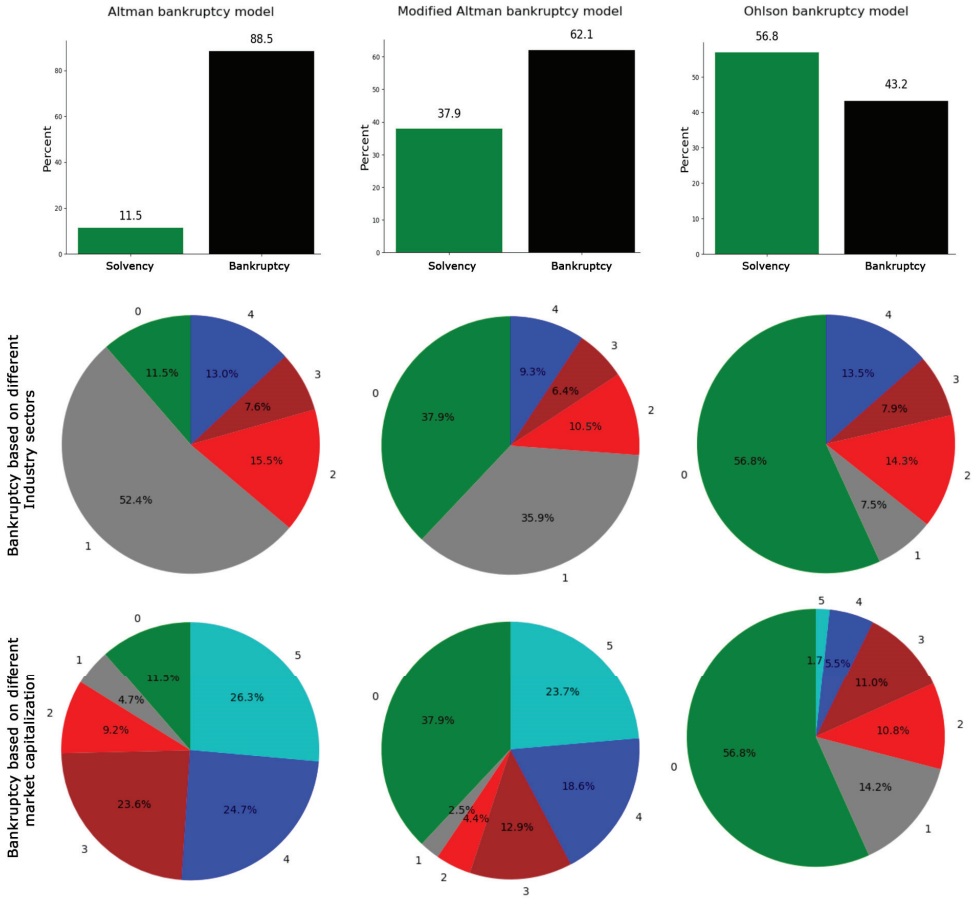


Figure 2. Percentage difference insolvency/bankruptcy prediction across various bankruptcy prediction models. Pie charts display the differences in bankruptcy and solvency percentages based on various industry subsectors and market capitalization classifications using three bankruptcy models. Industry subsector classes are labeled as 0 for solvency and from 1 to 4 for bankruptcy in biotechnology, medical devices, pharmacology, and healthcare services, respectively. Market capitalization classes are defined as 0 for solvency and from 1 to 5 for bankruptcy in large-cap, mid-cap, small-cap, micro-cap, and nano-cap categories, respectively.

Figure 3 presents a heatmap of the confusion matrix for our final tuned gradient boosting model, offering a visual representation of the model's performance across different classes. The gradation of colors reflects the proportion of true positives, true negatives, false positives, and false negatives, thereby providing an intuitive overview of classification accuracy and misclassification patterns. The heatmap of the confusion matrix, as depicted in Figure 3, further corroborates the high accuracy and robust performance of the gradient boosting model in predicting bankruptcy based on financial ratios. The predominance

of high values along the matrix's main diagonal and lower values in off-diagonal cells indicates a substantial rate of correct predictions relative to the misclassifications. This visualization underscores the gradient boosting model's effectiveness in distinguishing between the various classes.

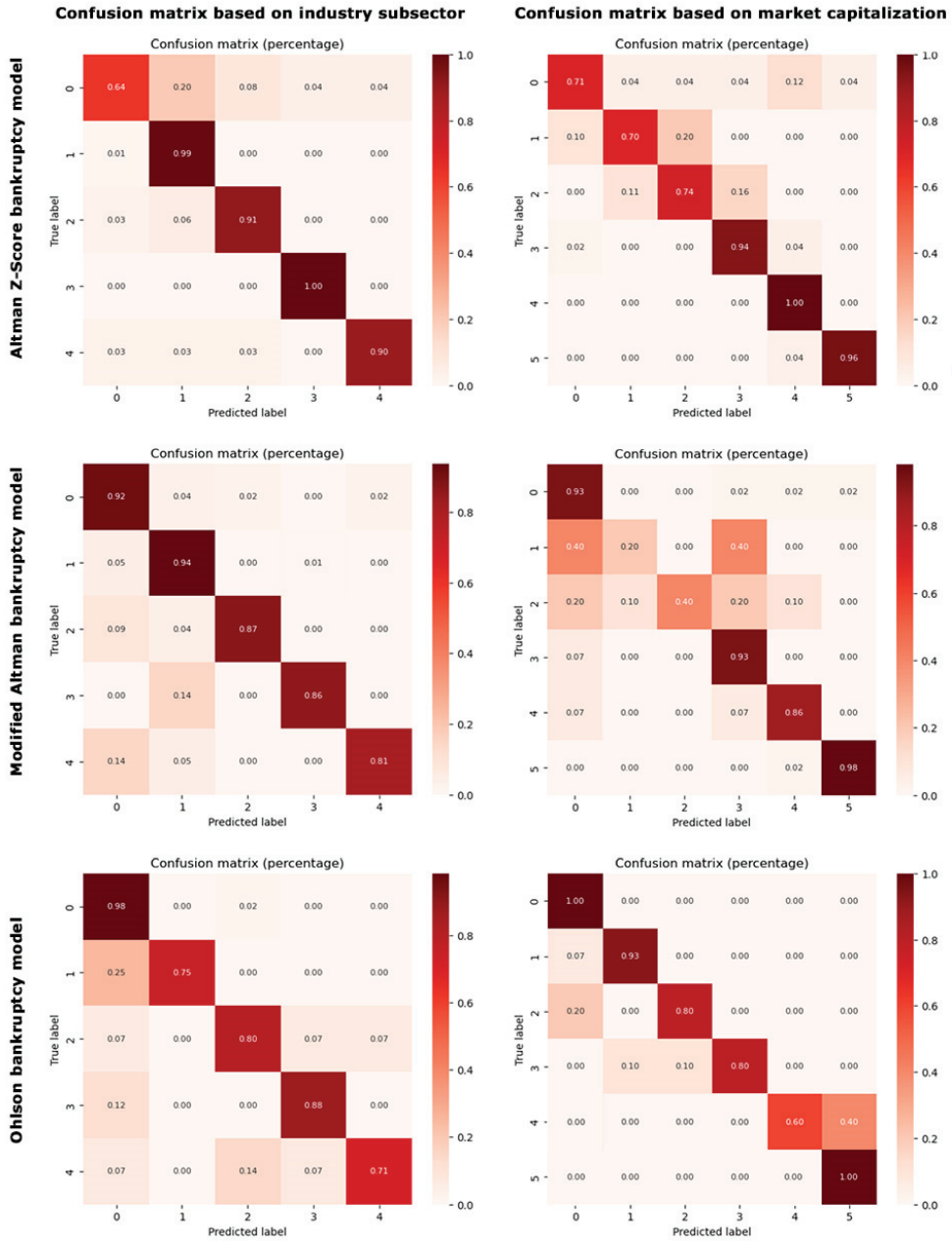


Figure 3. Heatmaps of the confusion matrices showcasing the classification models performance in differentiating between solvency and bankruptcy across industry subsectors and market capitalization classes. Industry subsector classes are labeled as 0 for solvency and from 1 to 4 for bankruptcy in biotechnology, medical devices, pharmacology, and healthcare services, respectively. Market capitalization classes are defined as 0 for solvency and from 1 to 5 for bankruptcy in large-cap, mid-cap, small-cap, micro-cap, and nano-cap categories, respectively.

Given the critical importance of minimizing false negatives in bankruptcy prediction, where failing to identify an impending bankruptcy could have drastic financial repercussions, the receiver operating characteristic (ROC) curve's ability to illustrate the trade-off between sensitivity (true-positive rate) and specificity (false-positive rate) across different thresholds makes it a fitting choice. It facilitates the identification of an optimal balance,

ensuring that the model minimally overlooks actual cases of bankruptcy, which is of paramount concern in financial analytics. Therefore, we compared ROC curves in Figure 4 based on different bankruptcy models and classifications for the gradient boosting model. The areas under the ROC curves for the Altman Z-score, modified Altman, and Ohlson bankruptcy prediction models were notably high (greater than 90%) for classifications based on industry subsectors and market capitalization. The high ROC curve areas show the exceptional capability of our gradient boosting algorithm in differentiating bankrupt from solvent healthcare industries. This suggests not only the accuracy of our predictions but also confirms their consistency and robustness across diverse industry contexts and varying scales of market capitalization. The superior performance of the gradient boosting model and the suitable size of our dataset reinforce the notion that this efficient and transparent model is advantageous, particularly in scenarios where complexity is unnecessary, and high-dimensional deep learning approaches do not confer additional benefits.

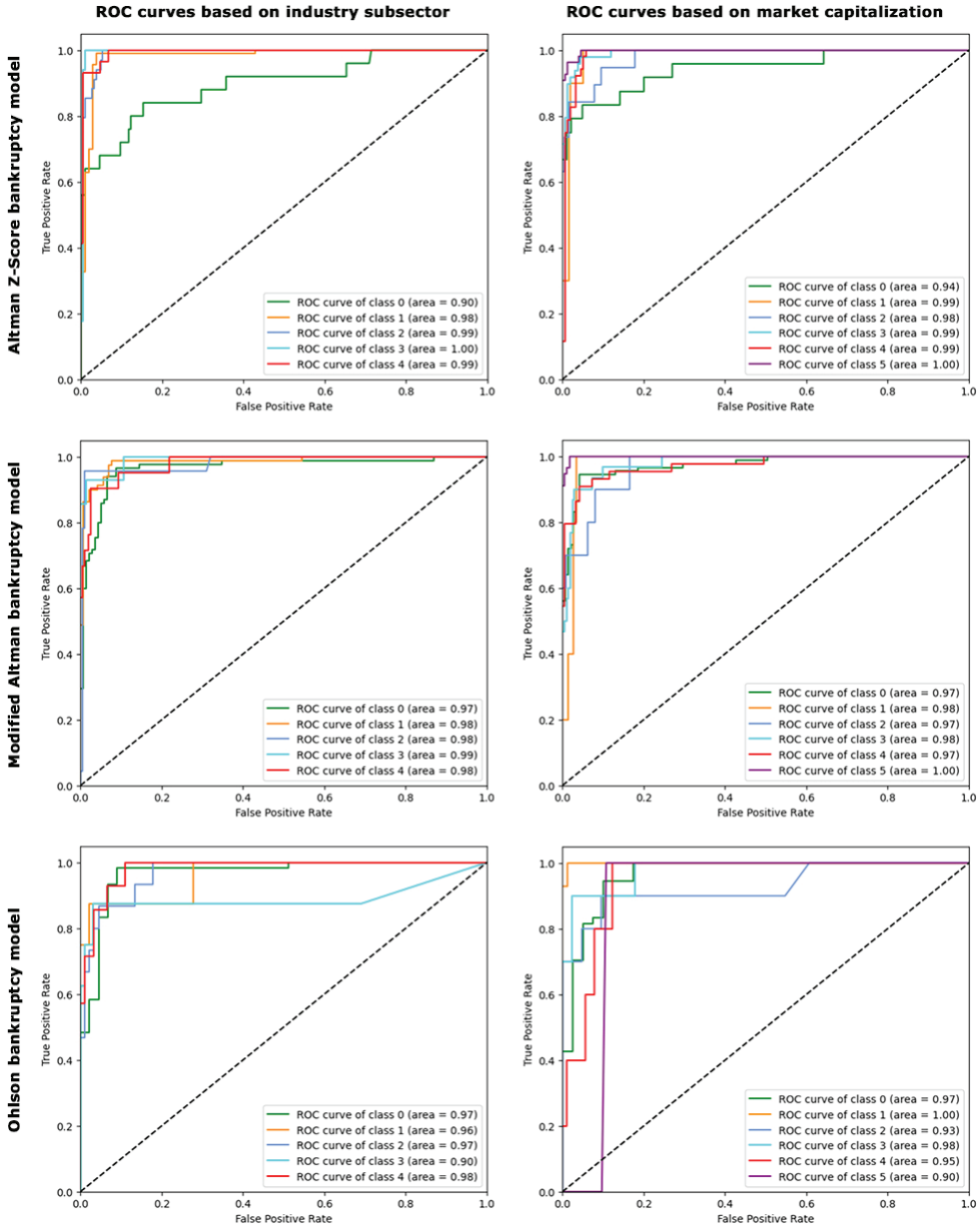


Figure 4. The ROC curves illustrating the true-positive rate against the false-positive rate at various threshold settings for the classification models applied to industry subsectors and market capitalization classes. ROC, receiver operating characteristic.

6. Discussion

Predicting the probability of bankruptcy remains among the most formidable challenges in contemporary economic and financial research [1,2]. Specifically, the evaluation of bankruptcy risk in the industries and, more precisely, within the medical and healthcare sectors holds immense significance due to its profound financial implications. Consequently, the present study was designed to achieve a precise prediction of bankruptcy risk for U.S. healthcare industries based on comprehensive financial ratios that encapsulate a wide array of financial metrics. This allows for the early detection of financial fragilities, the protection of economic interests, and the provision of timely insights for stakeholders. Effective prediction aids in the enhancement of management strategies, the attainment of well-informed advancement and risk control measures, and ensures the provision of high-quality healthcare services to not only the American population but potentially the global community as well. Our results indicated high bankruptcy risks as predicted by Altman, modified Altman, and Ohlson bankruptcy models, which can be attributed to the distinctive economic landscape of 2022. This can be related to significant industry volatility driven by the COVID-19 pandemic, escalating inflationary pressures, and a series of aggressive interest rate hikes in 2022. These factors led to increased borrowing costs that impacted market liquidity and valuations, particularly within growth-oriented sectors such as healthcare, which are traditionally more sensitive to interest rate changes.

6.1. The High Robustness and Predictive Power of Gradient Boosting Machine Learning Algorithm

Logical and meaningful improvements in cross-validation accuracies following hyperparameter tuning (by 7.1–10.3%), combined with the remarkably high levels of post-tuning accuracy (exceeding 90%) for the gradient boosting model, as well as elevated precision, recall, and F1 scores—as evidenced by the confusion matrices—and areas under the ROC curve surpassing 0.9 jointly affirm the robustness and exceptional predictive capability of this model. Therefore, the gradient boosting model can efficiently predict the bankruptcy risk of healthcare industries based on financial ratios with high accuracy and robustness. The study by Carmona et al. on predicting financial distress in French firms corroborated and further supported our findings regarding the effectiveness of gradient boosting algorithms in financial distress prediction [36,37]. The results also revealed that even the accuracies of our other machine learning algorithms were higher than those reported in some corresponding studies related to the prediction of bankruptcy risk based on other financial parameters, even with a larger database [60]. This underscores the power and sensitivity of financial ratios in predicting bankruptcy within the healthcare industries, providing creditors, investors, and other stakeholders with the opportunity to take proactive measures to mitigate financial challenges.

6.2. Important Financial Ratios Sensitive to Bankruptcy Prediction

The feature importance analysis presented in Figure 5 indicated that, according to both Altman models, the most important and sensitive ratio for predicting bankruptcy risk across various industry subsectors and market capitalizations is the return on investment (ROI). ROI is a measure comparing net income to investment. A high ROI indicates that the investment's gains are favorable compared to its cost. Conversely, the Ohlson model identified the return on assets (ROA) as the important ratio in predicting bankruptcy risk specific to industry subsectors, with a feature importance of 7.8%. ROA is a financial metric showing a company's profitability relative to its total assets. It helps determine if a company efficiently uses its assets to generate profit. For the predictions based on market capitalization, the Ohlson model highlighted not only the ROA ratio with an importance of 6.3% but also the enterprise value to earnings before interest and taxes (EV/EBIT) ratio, which has an importance of 6.6%. The comprehensive details regarding the relative importance of all parameters under investigation are illustrated in Figure 5. Notwithstanding the absence of prior scholarly work specific to bankruptcy risk prediction within the healthcare industries, the extant literature on bankruptcy assessment in disparate

industries also reinforces the significance of several ratios analogous to those we identified as pivotal [12–15,61].

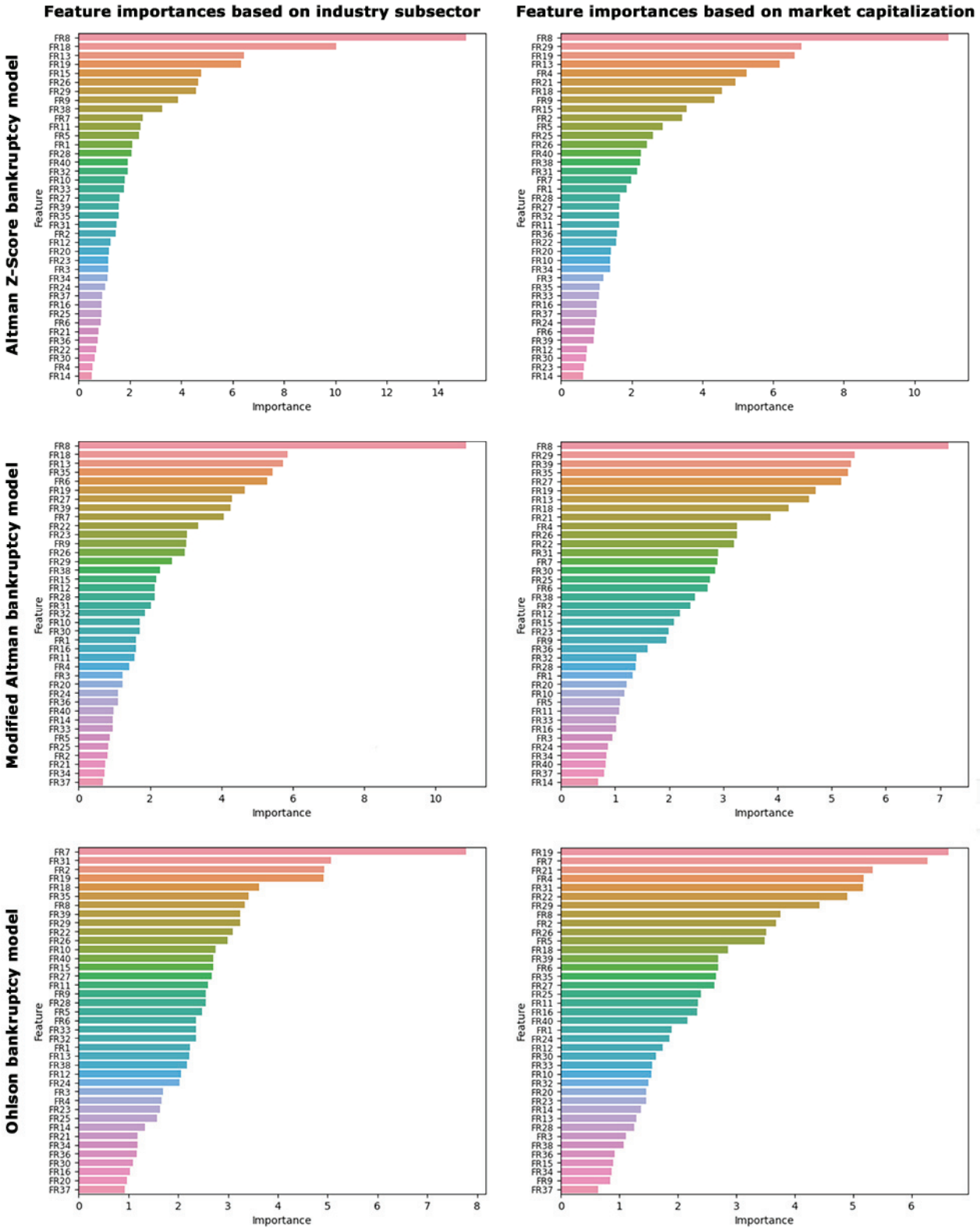


Figure 5. Feature importance analysis highlighting the relative importance of each financial ratio in determining bankruptcy risk based on different industry subsectors and market capitalizations.

The importance and sensitivity of the ROI as illuminated by the Altman models suggest that investors and analysts in medical and healthcare sectors place considerable emphasis on the immediate efficiency of capital usage. This ratio's predictive potency could potentially stem from its ability to encapsulate the sector's unique capital structures and reinvestment patterns critical to sustaining operations in the complex healthcare market. On the other hand, the Ohlson model's emphasis on ROA accentuates the critical role of effective asset utilization in the financial viability of healthcare firms, hinting at a market environment where operational efficiency is a harbinger of fiscal health. Additionally, the Ohlson model's accentuation of the EV/EBIT ratio underscores the market's valuation of profitability relative to a company's total valuation. The nuanced divergence observed between the models regarding the most influential ratios offers compelling insights into the multifaceted nature of bankruptcy risk assessment in this sector, underscoring the need for a multidimensional analytic approach in financial prognostication efforts within the healthcare industries. Consequently, for an augmented and holistic forecast of bankruptcy risk, it is incumbent upon both researchers and practitioners to amalgamate the varied lenses provided by the Altman and Ohlson models.

It should be noted that our investigation, as described, revealed widespread financial distress among U.S. healthcare industries in 2022. This trend might partly be attributed to these key financial ratios identified by these models, which likely reflect the broader fiscal challenges faced by these entities. These bankruptcy models were originally conceptualized to be effective over a two-year period; hence, the important and sensitive parameters suggested by these models theoretically remain relevant and effective until 2024. Financial analysts should continue to monitor these ratios closely to maintain the solvency of the medical and healthcare sectors. Therefore, recognizing these ratios not only offers a retrospective analysis of past financial difficulties but also provides strategic foresight. By vigilantly tracking these financial indicators, stakeholders can better prepare to address fiscal vulnerabilities proactively, thereby reducing the risk of similar financial challenges in the future. In addition to managing the aforementioned financial ratios based on our findings, recent advancements in medical technology also help address some challenges in the financial aspects of healthcare industries [62–71]. Therefore, technological advancements can be beneficial beyond just controlling our recognized financial ratios.

6.3. Limitations and Future Directions

It should be noted that, although 40 financial ratios covering key financial domains were utilized to predict bankruptcy, future research could enhance our findings by incorporating additional predictive parameters. Alzayed et al. recently highlighted the importance of corporate governance alongside financial ratios in the context of bank failure predictions [72]. Hence, incorporating factors like corporate governance, competitive dynamics, and regulatory considerations may also prove significant in enhancing the prediction of bankruptcy for the healthcare industries, an area ripe for future academic exploration. Future studies could benefit from repeating our analysis using data from a more typical financial year as a control group. This comparison would allow researchers to discern whether the financial metrics identified as crucial during the crisis conditions of 2022 hold similar significance in more stable times. Such a study would also help in evaluating the robustness of these metrics across varying economic conditions, providing a deeper understanding of their utility in continuous risk management and strategic financial planning in the healthcare sector. Advancements in natural language processing in future studies could provide mechanisms to evaluate market sentiment and management discourse, while machine learning models like the gradient boosting algorithm demonstrated in this study could be further refined by integrating sector-specific data for a more tailored risk assessment. Exploring the impact of emergent global health issues and economic disruptions beyond COVID-19 may also be essential in shaping more resilient financial prediction frameworks for the sector.

7. Conclusions

This study predicted the risk of bankruptcy in the U.S. healthcare industries based on financial ratios across different industry sub-sectors and varying market capitalizations, using machine learning analysis. The results suggest that financial ratios serve as robust predictors of bankruptcy, and the gradient boosting algorithm can significantly enhance the predictive power of conventional models. This study has elucidated important financial ratios that served as potential harbingers of financial distress and heightened bankruptcy risk within U.S. healthcare industries during 2022. The identification of these ratios is instrumental not only for signaling potential red flags but also for equipping stakeholders with the insights necessary to devise strategies aimed at risk mitigation and ensuring the sustainability of the healthcare industries. The comprehensive perspective offered through this research enriches the academic discourse on bankruptcy prediction and offers strategic foresight for stakeholders within healthcare industries. Ultimately, the findings lay a robust foundation for future research and offer a framework for informed decision making that serves investors, policymakers, shareholders, and healthcare professionals, thereby contributing significantly to the financial stability and resilience of the medical and healthcare sectors.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jtaer19020066/s1, Supplementary Material S1: The definitions of the 40 financial ratios (features) used to assess the bankruptcy risk of medical and healthcare stocks in the U.S.; FR, financial ratio; FR, financial ratio. Supplementary Material S2: Overview of multi-class classification techniques that are used in the present study. Supplementary Material S3: Performance metrics and hyperparameters of bankruptcy prediction models for different machine and ensemble learning classifiers.

Author Contributions: H.G. contributed to conceptualization, study design, data collection, data pre-processing, investigation, methodology, machine learning analysis, software development, validation, and manuscript writing; M.A. contributed to data collection, investigation, project administration, data pre-processing, and initial draft preparation. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the College of Business of Northern Michigan University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data can bot be shared from the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Clement, C. Machine Learning in Bankruptcy Prediction—A Review. *J. Public Adm. Financ. Law* **2020**, *17*, 178–196.
2. Geng, R.; Bose, I.; Chen, X. Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *Eur. J. Oper. Res.* **2015**, *241*, 236–247. [CrossRef]
3. Wessels, M. Predicting Financial Distress at Dutch General Hospitals: A Machine Learning Approach. Master's Thesis, Utrecht University, Utrecht, The Netherlands, 2018.
4. Beauvais, B.; Ramamonjiarivelo, Z.; Betancourt, J.; Cruz, J.; Fulton, L. The Predictive Factors of Hospital Bankruptcy—An Exploratory Study. *Healthcare* **2023**, *11*, 165. [CrossRef] [PubMed]
5. Pendergast, P.M.; Sousa, M.D.; Wadsworth, T. Health Insurance and Bankruptcy Risk: Examining the Impact of the Affordable Care Act. *Brook. L. Rev.* **2020**, *86*, 975.
6. Jones, B.D. The Government's Perspective on Health Care Bankruptcies. *Am. Bankruptcy Inst. J.* **2018**, *37*, 22–23.
7. Skinner, B.J. *The Medical Bankruptcy Myth*; Fraser Institute: Vancouver, BC, Canada, 2009.
8. Shrime, M.G.; Weinstein, M.C.; Hammitt, J.K.; Cohen, J.L.; Salomon, J.A. Trading bankruptcy for health: A discrete-choice experiment. *Value Health* **2018**, *21*, 95–104. [CrossRef] [PubMed]
9. Himmelstein, D.U.; Thorne, D.; Warren, E.; Woolhandler, S. Medical bankruptcy in the United States, 2007: Results of a national study. *Am. J. Med.* **2009**, *122*, 741–746. [CrossRef]
10. Gross, T.; Notowidigdo, M.J. Health insurance and the consumer bankruptcy decision: Evidence from expansions of Medicaid. *J. Public Econ.* **2011**, *95*, 767–778. [CrossRef]

11. Maizel, S.; Bernardino, C.; Caine, M.; Garfinkle, J. Corporate Bankruptcy Panel: The Healthcare Industry Post-Affordable Care Act: A Bankruptcy Perspective. *Emory Bankr. Dev. J.* **2014**, *31*, 249.
12. Supriyanto, J.; Darmawan, A. The effect of financial ratio on financial distress in predicting bankruptcy. *J. Appl. Manag. Account.* **2018**, *2*, 110–120. [CrossRef]
13. Amalia, S.; Fadjriah, N.E.; Nugraha, N.M. The influence of the financial ratio to the prevention of bankruptcy in cigarette manufacturing companies sub sector. *Solid State Technol.* **2020**, *63*, 4173–4182.
14. Lee, B.-H.; Lee, S.-H. A study on financial ratio and prediction of financial distress in financial markets. *J. Distrib. Sci.* **2018**, *16*, 21–27.
15. Tian, S.; Yu, Y. Financial ratios and bankruptcy predictions: An international evidence. *Int. Rev. Econ. Financ.* **2017**, *51*, 510–526. [CrossRef]
16. Aly, S.; Alfonse, M.; Salem, A.-B.M. Bankruptcy prediction using artificial intelligence techniques: A survey. In *Digital Transformation Technology: Proceedings of ITAF 2020, Online, 26–27 January 2021*; Springer: Singapore, 2022; pp. 335–360.
17. Sarker, I.H. Machine learning: Algorithms, real-world applications and research directions. *SN Comput. Sci.* **2021**, *2*, 160. [CrossRef] [PubMed]
18. Rao, T.V.N.; Gaddam, A.; Kurni, M.; Saritha, K. Reliance on artificial intelligence, machine learning and deep learning in the era of industry 4.0. In *Smart Healthcare System Design: Security and Privacy Aspects*; John Wiley & Sons: Hoboken, NJ, USA, 2022; pp. 281–299.
19. Huang, Y.-P.; Yen, M.-F. A new perspective of performance comparison among machine learning algorithms for financial distress prediction. *Appl. Soft Comput.* **2019**, *83*, 105663. [CrossRef]
20. Roumani, Y.F.; Nwankpa, J.K.; Tanniru, M. Predicting firm failure in the software industry. *Artif. Intell. Rev.* **2020**, *53*, 4161–4182. [CrossRef]
21. Smith, M.; Alvarez, F. Predicting firm-level bankruptcy in the Spanish economy using extreme gradient boosting. *Comput. Econ.* **2022**, *59*, 263–295. [CrossRef]
22. Zelenkov, Y.; Fedorova, E.; Chekrizov, D. Two-step classification method based on genetic algorithm for bankruptcy forecasting. *Expert Syst. Appl.* **2017**, *88*, 393–401. [CrossRef]
23. Schönfeld, J.; Kuděj, M.; Smrčka, L. Financial health of enterprises introducing safeguard procedure based on bankruptcy models. *J. Bus. Econ. Manag.* **2018**, *19*, 692–705. [CrossRef]
24. Tsai, C.-F.; Hsu, Y.-F.; Yen, D.C. A comparative study of classifier ensembles for bankruptcy prediction. *Appl. Soft Comput.* **2014**, *24*, 977–984. [CrossRef]
25. Alfaro, E.; García, N.; Gámez, M.; Elizondo, D. Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks. *Decis. Support Syst.* **2008**, *45*, 110–122. [CrossRef]
26. Mai, F.; Tian, S.; Lee, C.; Ma, L. Deep learning models for bankruptcy prediction using textual disclosures. *Eur. J. Oper. Res.* **2019**, *274*, 743–758. [CrossRef]
27. Antulov-Fantulin, N.; Lagravinese, R.; Resce, G. Predicting bankruptcy of local government: A machine learning approach. *J. Econ. Behav. Organ.* **2021**, *183*, 681–699. [CrossRef]
28. Shetty, S.; Musa, M.; Brédart, X. Bankruptcy Prediction Using Machine Learning Techniques. *J. Risk Financ. Manag.* **2022**, *15*, 35. [CrossRef]
29. Lahmiri, S.; Bekiros, S. Can machine learning approaches predict corporate bankruptcy? Evidence from a qualitative experimental design. *Quant. Financ.* **2019**, *19*, 1569–1577. [CrossRef]
30. Kim, H.; Cho, H.; Ryu, D. Corporate bankruptcy prediction using machine learning methodologies with a focus on sequential data. *Comput. Econ.* **2022**, *59*, 1231–1249. [CrossRef]
31. Liashenko, O.; Kravets, T.; Kostovetskyi, Y. Machine Learning and Data Balancing Methods for Bankruptcy Prediction. *Ekonomika* **2023**, *102*, 28–46. [CrossRef]
32. Jan, C.-L. Financial information asymmetry: Using deep learning algorithms to predict financial distress. *Symmetry* **2021**, *13*, 443. [CrossRef]
33. Zoričák, M.; Gnip, P.; Drotár, P.; Gazda, V. Bankruptcy prediction for small-and medium-sized companies using severely imbalanced datasets. *Econ. Model.* **2020**, *84*, 165–176. [CrossRef]
34. Adosoglou, G.; Lombardo, G.; Pardalos, P.M. Neural network embeddings on corporate annual filings for portfolio selection. *Expert Syst. Appl.* **2021**, *164*, 114053. [CrossRef]
35. Bazzana, F.; Bee, M.; Hussin Adam Khatir, A.A. Machine learning techniques for default prediction: An application to small Italian companies. *Risk Manag.* **2024**, *26*, 1. [CrossRef]
36. Carmona, P.; Dwekat, A.; Mardawi, Z. No more black boxes! Explaining the predictions of a machine learning XGBoost classifier algorithm in business failure. *Res. Int. Bus. Financ.* **2022**, *61*, 101649. [CrossRef]
37. Ben Jabeur, S.; Stef, N.; Carmona, P. Bankruptcy prediction using the XGBoost algorithm and variable importance feature engineering. *Comput. Econ.* **2023**, *61*, 715–741. [CrossRef]
38. Lombardo, G.; Pellegrino, M.; Adosoglou, G.; Cagnoni, S.; Pardalos, P.M.; Poggi, A. Machine Learning for Bankruptcy Prediction in the American Stock Market: Dataset and Benchmarks. *Future Internet* **2022**, *14*, 244. [CrossRef]
39. Adosoglou, G.; Park, S.; Lombardo, G.; Cagnoni, S.; Pardalos, P.M. Lazy network: A word embedding-based temporal financial network to avoid economic shocks in asset pricing models. *Complexity* **2022**, *2022*, 9430919. [CrossRef]

40. Bragoli, D.; Ferretti, C.; Ganugi, P.; Marseguerra, G.; Mezzogori, D.; Zammori, F. Machine-learning models for bankruptcy prediction: Do industrial variables matter? *Spat. Econ. Anal.* **2022**, *17*, 156–177. [CrossRef]
41. Liu, Y.; Zeng, Q.; Li, B.; Ma, L.; Ordieres-Meré, J. Anticipating financial distress of high-tech startups in the European Union: A machine learning approach for imbalanced samples. *J. Forecast.* **2022**, *41*, 1131–1155. [CrossRef]
42. Papík, M.; Papíková, L. Impacts of crisis on SME bankruptcy prediction models' performance. *Expert Syst. Appl.* **2023**, *214*, 119072.
43. Liu, Q.; Shi, C.; Tse, Y.; Zhang, L. The impact of public health emergencies on small and medium-sized enterprises: Evidence from China. *Glob. Financ. J.* **2023**, *58*, 100892. [CrossRef]
44. Tudose, M.B.; Avasilcai, S. A review of the research on financial performance and its determinants. In *Innovation in Sustainable Management and Entrepreneurship: Proceedings of the 2019 International Symposium in Management (SIM2019), Timisoara, Romania, 25–26 October 2019*; Springer: Cham, Switzerland, 2020; pp. 229–244.
45. Najib, A.S.; Cahyaningdyah, D. Analysis of the bankruptcy of companies with Altman model and Ohlson model. *Manag. Anal. J.* **2020**, *9*, 243–251. [CrossRef]
46. Sharma, S.; Bodla, B.S. Review and comparison of Altman and Ohlson model to predict bankruptcy of companies. *ANVESAK* **2022**, *52*, 30–36.
47. Altman, E.I.; Iwanicz-Drozdowska, M.; Laitinen, E.K.; Suvas, A. Distressed firm and bankruptcy prediction in an international context: A review and empirical analysis of Altman's Z-score model. *SSRN Electron. J.* **2014**, 2536340. [CrossRef]
48. Elviani, S.; Simbolon, R.; Riana, Z.; Khairani, F.; Dewi, S.P.; Fauzi, F. The Accuracy of the Altman, Ohlson, Springate and Zmejewski Models in Bankruptcy Predicting Trade Sector Companies in Indonesia. *Bp. Int. Res. Crit. Inst. (BIRCI-J.)* **2020**, *3*, 334–347. [CrossRef]
49. Salimi, A.Y. Validity of Altmans Z-Score model in predicting Bankruptcy in recent years. *Acad. Account. Financ. Stud. J.* **2015**, *19*, 233.
50. Ohlson, J.A. Financial ratios and the probabilistic prediction of bankruptcy. *J. Account. Res.* **1980**, *18*, 109–131. [CrossRef]
51. Altman, E.I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Financ.* **1968**, *23*, 589–609. [CrossRef]
52. Altman, E.I.; Iwanicz-Drozdowska, M.; Laitinen, E.K.; Suvas, A. Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-score model. *J. Int. Financ. Manag. Account.* **2017**, *28*, 131–171. [CrossRef]
53. Matenda, F.R.; Sibanda, M.; Chikodza, E.; Gumbo, V. Bankruptcy prediction for private firms in developing economies: A scoping review and guidance for future research. *Manag. Rev. Q.* **2022**, *72*, 927–966. [CrossRef]
54. Armeanu, D.-S.; Cioaca, S.-I. An assessment of the bankruptcy risk on the Romanian capital market. *Procedia-Soc. Behav. Sci.* **2015**, *182*, 535–542. [CrossRef]
55. Frino, A.; Jones, S.; Wong, J.B. Market behaviour around bankruptcy announcements: Evidence from the Australian Stock Exchange. *Account. Financ.* **2007**, *47*, 713–730. [CrossRef]
56. Fauzi, R.; Wahyudi, I. The effect of firm and stock characteristics on stock returns: Stock market crash analysis. *J. Financ. Data Sci.* **2016**, *2*, 112–124. [CrossRef]
57. Deng, P.L. The Effect of Trading Activity and Holdings Market Capitalization on Portfolio Performance. *Int. J. Econ. Financ.* **2018**, *10*, 1–18. [CrossRef]
58. Gholampour, S. Impact of Nature of Medical Data on Machine and Deep Learning for Imbalanced Datasets: Clinical Validity of SMOTE Is Questionable. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 827–841. [CrossRef]
59. Waterstraat, M.G.; Dehghan, A.; Gholampour, S. Optimization of number and range of shunt valve performance levels in infant hydrocephalus: A machine learning analysis. *Front. Bioeng. Biotechnol.* **2024**, *12*, 1352490. [CrossRef] [PubMed]
60. Barboza, F.; Kimura, H.; Altman, E. Machine learning models and bankruptcy prediction. *Expert Syst. Appl.* **2017**, *83*, 405–417. [CrossRef]
61. Singh, S.G. The early indicators of financial failure: A study of bankrupt and solvent health systems. *J. Healthc. Manag.* **2008**, *53*, 333.
62. Gholampour, S. Computerized biomechanical simulation of cerebrospinal fluid hydrodynamics: Challenges and opportunities. *Comput. Methods Programs Biomed.* **2021**, *200*, 105938. [CrossRef] [PubMed]
63. Gholampour, S.; Deh, H.H.H. The effect of spatial distances between holes and time delays between bone drillings based on examination of heat accumulation and risk of bone thermal necrosis. *Biomed. Eng. Online* **2019**, *18*, 65. [CrossRef]
64. Gholampour, S.; Droessler, J.; Frim, D. The role of operating variables in improving the performance of skull base grinding. *Neurosurg. Rev.* **2022**, *45*, 2431–2440. [CrossRef]
65. Gholampour, S.; Hassanalideh, H.H.; Gholampour, M.; Frim, D. Thermal and physical damage in skull base drilling using gas cooling modes: FEM simulation and experimental evaluation. *Comput. Methods Programs Biomed.* **2021**, *212*, 106463. [CrossRef] [PubMed]
66. Hosseini, S.S.; Yamini, B.; Ichkitidze, L.; Asadi, M.; Fernandez, J.; Gholampour, S. Enhanced ionic polymer–metal composites with nanocomposite electrodes for restoring eyelid movement of patients with ptosis. *Nanomaterials* **2023**, *13*, 473. [CrossRef]
67. Hassanalideh, H.H.; Gholampour, S. Finding the optimal drill bit material and proper drilling condition for utilization in the programming of robot-assisted drilling of bone. *CIRP J. Manuf. Sci. Technol.* **2020**, *31*, 34–47. [CrossRef]
68. Gholampour, S.; Hajirayat, K. Minimizing thermal damage to vascular nerves while drilling of calcified plaque. *BMC Res. Notes* **2019**, *12*, 338. [CrossRef] [PubMed]

69. Gholampour, S. Feasibility of assessing non-invasive intracranial compliance using FSI simulation-based and MR elastography-based brain stiffness. *Sci. Rep.* **2024**, *14*, 6493. [CrossRef] [PubMed]
70. Gholampour, S. Can magnetic resonance elastography serve as a diagnostic tool for gradual-onset brain disorders? *Neurosurg. Rev.* **2023**, *47*, 3. [CrossRef] [PubMed]
71. Gholampour, S. Modeling and simulation of cerebrospinal fluid disorders. *Front. Bioeng. Biotechnol.* **2023**, *11*, 1331170. [CrossRef]
72. Alzayed, N.; Eskandari, R.; Yazdifar, H. Bank failure prediction: Corporate governance and financial indicators. *Rev. Quant. Financ. Account.* **2023**, *61*, 601–631. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Journal of
Theoretical and Applied
Electronic Commerce Research

MDPI

Article

Quantitative Stock Selection Model Using Graph Learning and a Spatial–Temporal Encoder

Tianyi Cao [1,2], Xinrui Wan [1], Huanhuan Wang [1], Xin Yu [1] and Libo Xu [1,*]

[1] School of Computing and Data Engineering, Ningbo Tech University, Ningbo 315000, China; 2307728636@nit.zju.edu.cn (T.C.); wanxinrui@nbt.edu.cn (X.W.); whh@nit.zju.edu.cn (H.W.); yuxin@zju.edu.cn (X.Y.)
[2] College of Business and Economics, Australian National University, Canberra 0200, Australia
* Correspondence: xlb@nit.zju.edu.cn

Abstract: In the rapidly evolving domain of finance, quantitative stock selection strategies have gained prominence, driven by the pursuit of maximizing returns while mitigating risks through sophisticated data analysis and algorithmic models. Yet, prevailing models frequently neglect the fluid dynamics of asset relationships and market shifts, a gap that undermines their predictive and risk management efficacy. This oversight renders them vulnerable to market volatility, adversely affecting investment decision quality and return consistency. Addressing this critical gap, our study proposes the Graph Learning Spatial–Temporal Encoder Network (GL-STN), a pioneering model that seamlessly integrates graph theory and spatial–temporal encoding to navigate the intricacies and variabilities of financial markets. By harnessing the inherent structural knowledge of stock markets, the GL-STN model adeptly captures the nonlinear interactions and temporal shifts among assets. Our innovative approach amalgamates graph convolutional layers, attention mechanisms, and long short-term memory (LSTM) networks, offering a comprehensive analysis of spatial–temporal data features. This integration not only deciphers complex stock market interdependencies but also accentuates crucial market insights, enabling the model to forecast market trends with heightened precision. Rigorous evaluations across diverse market boards—Main Board, SME Board, STAR Market, and ChiNext—underscore the GL-STN model's exceptional ability to withstand market turbulence and enhance profitability, affirming its substantial utility in quantitative stock selection.

Keywords: GL-STN; quantitative stock selection; spatial–temporal encoder; graph convolution; graph learning

1. Introduction

Quantitative stock selection uses knowledge from statistics, computer science, and finance to create models for selecting stock portfolios. The goal is to achieve returns that exceed benchmark rates. With the development of big data and cloud computing technologies, financial time series data exhibit more complex characteristics, including high dimensionality, strong volatility, and a large scale. Consequently, quantitative investment techniques are evolving from traditional strategies to modern artificial intelligence technologies, such as machine learning and deep learning.

Quantitative stock selection methods based on modern AI can be broadly categorized into machine learning and deep learning approaches. Machine learning methods, including random forest [1], gradient boosting [2], and support vector machines (SVMs) [3], are trained on features and historical statistical characteristics without considering data temporality. These models face performance limitations with large-scale, unstructured, or high-dimensional data. They struggle to capture the complex, nonlinear patterns in stock markets. Deep learning methods, represented by recurrent neural networks (RNN) [4] and long short-term memory (LSTM) networks, can process sequential data directly [5].

Compared to traditional linear models, these deep learning approaches offer significant improvements. They are more effective in predicting stock price movements, identifying market opportunities, and constructing automated trading systems. Transformers, known for handling long-term dependencies in sequential data, have emerged as another focal point for time series forecasting. Some researchers have explored hybrid models that combine neural networks with statistical methods to enhance financial forecasting accuracy [6]. Despite the progress made in deep learning-based quantitative stock selection, challenges still remain due to the high dimensionality, volatility [7], and complex interrelations of financial time series. Most existing models handle stocks or assets in isolation, overlooking potential correlations and interactions. This results in the lack of a relational modeling perspective. Recent studies leveraging graph neural networks have sought to address this issue, but reliance on subjective graph construction may distort input information. Also, the lack of focus on the importance of individual stocks reduces the accuracy of graph node representations. Moreover, integrating relational modeling with the sequential nature of time series data remains an area for improvement.

To tackle these challenges, we propose the Graph Learning Spatial–Temporal Encoder Network (GL-STN) model. This model effectively addresses the high dimensionality, strong volatility, and complex correlations of financial time series data by using graph embedding learning, modeling temporal and spatial relationships.

The GL-STN model overcomes the limitations of existing models by introducing innovative methods for graph learning and spatial–temporal encoding. It effectively addresses the high dimensionality and volatility of financial data and captures the complex interrelations between stocks.

Unlike traditional machine learning models that struggle with large-scale, unstructured, or high-dimensional data, GL-STN excels by modeling the intricate dependencies and temporal dynamics of financial markets. While conventional neural network models focus on sequential data processing, GL-STN integrates graph convolutional neural networks with LSTM to capture both spatial relationships and temporal dependencies. This dual capability allows for a more comprehensive analysis of market data. Moreover, compared to existing graph neural network models, GL-STN reduces subjective factors in the graph construction process. It accurately determines the importance of each stock node through attention mechanisms, enhancing the accuracy of node representations. This multi-faceted approach ensures unprecedented accuracy and robustness in financial market analysis.

Furthermore, the GL-STN model's multi-layered and multi-module architecture enables processing and analyzing data from multiple perspectives and dimensions. This facilitates comprehensive multi-modal feature extraction and processing while maintaining flexibility and scalability. Consequently, the GL-STN model can easily extend and adapt to incorporate additional data sources or new analytical techniques. This flexibility results in unprecedented accuracy and stability.

Key contributions include constructing a network graph that learns and establishes relationships between stocks, effectively capturing their interactions and dependencies. This approach addresses the oversight of stock correlations in traditional models. Additionally, graph convolutional neural networks and attention mechanisms [8] process graph structural information, extracting features and the importance of individual stocks for more accurate feature aggregation. Furthermore, by integrating LSTM, the model enhances its ability to capture temporal dependencies, facilitating the learning of features across time and space. Comparative analyses with mainstream machine learning and neural network models across various financial markets validate the effectiveness of the proposed model.

2. Related Work

2.1. Quantitative Investment Techniques Based on Machine Learning

Machine learning methods excel in model transparency and interpretability. They can demonstrate high efficiency and reliability when dealing with small-scale or structured

data. However, these methods often rely on feature engineering. They struggle to identify and utilize complex patterns and relationships hidden within the data. Wang S. [9] showcased the significant effectiveness of the random forest model in the stock market. Li B. et al. [10] used various machine learning models, including multiple neural networks, to construct different investment portfolios. Empirical tests showed annual returns significantly outperforming the market index. Zhou L. [11] confirmed the superiority of multi-factor stock selection models over traditional linear regression models through the study of six major categories of factors. Shu S. [12] created SCAD-logistic regression and MCP-logistic regression models by altering the penalty function in the logistic regression model, showing notable application effects in the stock market. Lewellen [13] successfully predicted excess returns of stocks by integrating 15 factors using the FM regression method. DeMiguel et al. [14] approached this topic from the perspective of investor utility, finding six company characteristics that could independently predict average returns. The utility-based models were practically applied in stock market prediction. Gu et al. [15] tested common machine learning algorithms in the U.S. market, finding that machine learning models could effectively outperform traditional linear regression models. Yu et al. [16] developed a support vector machine model to predict stock market trends. By using genetic algorithms to select attributes, the complexity of support vector machines can be reduced while improving modeling efficiency. In terms of classification performance, it can surpass traditional time series and neural network models. Based on intelligent media and technical analysis, Mndawe et al. [17] developed a stock price prediction framework, focusing on the application of machine learning in stock market prediction.

2.2. Quantitative Investment Techniques Based on Deep Neural Networks

By learning nonlinear changes in high-dimensional features, deep neural networks offer better representational capabilities than traditional methods. They can provide powerful analytical tools for quantitative investment. Despite challenges in model interpretability, the performance in prediction accuracy and efficiency has garnered widespread attention. Recent studies have shown significant progress in applying deep learning technology to quantitative stock selection strategies. Lin J. and Gong Z. [18] built convolutional neural network and BP neural network models based on the prices of Shanghai zinc futures, emphasizing the effectiveness of neural networks in stock price prediction. Liu et al. [19] proposed a quantitative option trading strategy based on convolutional neural networks. It is possible to exceed benchmark indices and basic momentum strategies by combining CNN predictions of stock prices with LSTM's quantitative timing strategy. Liang et al. [20] used long short-term memory networks to extract time series features and predict the trend of rebar futures prices. They found that using LSTM for feature extraction and the XGBoost model to filter trading signals generated by trading strategies can effectively improve annualized returns. Kim et al. [21] proposed a paired trading method combining deep reinforcement learning, enhancing the efficiency and profitability of trading strategies. Staffini [22] utilized the application of deep convolutional generative adversarial networks in stock price prediction, exploring the application of generative adversarial networks in financial market analysis.

The introduction of an attention mechanism provides new breakthroughs in the application of deep learning in quantitative investment strategies. Liu H et al. [23] explored how to use attention mechanisms to enhance the efficiency and effectiveness of quantitative investment strategies. By combining sequential graph structures with attention mechanisms, Lai et al. [24] developed new deep learning modules aimed at capturing and analyzing time series data more accurately.

2.3. Quantitative Investment Techniques Based on Graph Models

By constructing complex networks among financial assets, graph models [25] reveal the structure and dynamic changes in the market. They can provide a unique perspective for quantitative investment. In graph models, each asset in the financial market is treated

as a node, and interactions between assets are represented by edges within the network. This model is particularly suitable for analyzing complex systems in financial markets, revealing potential associations and mutual influences between assets. Wang Z. et al. [26] aimed to address complex relationships and data challenges in financial markets by classifying financial graphs and outlining graph neural network (GNN) methods. In order to propose a continuous learning factor predictor based on graph neural networks, Tan Z. et al. [27] focused on the challenges of future price prediction in high-frequency trading. By combining multi-factor pricing theory with real-time market dynamics, the limitations of existing methods in ignoring trend signals and interactions can be effectively overcome.

3. GL-STN

3.1. Model Framework

The framework of the GL-STN model, as shown in Figure 1, is divided into three modules: graph embedding, a graph convolutional network (GCN), and RNN, presenting a sequential relationship. The input sequence data represent the dynamic features of stocks. The nodes corresponding to each sequence are assigned a feature vector to represent the state of each strand. Initially, the feature vectors are iteratively optimized to compute the differences between nodes to construct an adjacency matrix. This step allows the model to utilize the graph's representational capability to map the correlations between individual stocks. Subsequently, the adjacency matrix and feature vectors are input into the GCN. Within the GCN, an attention mechanism is integrated to learn the importance distribution of different nodes across stocks, enhancing the model's precision in quantitative stock selection. Finally, the output from the GCN is passed to the LSTM to capture the temporal evolution patterns of stock features. By combining multiple layers and mechanisms, GL-STN generates outputs that provide observable support for subsequent quantitative stock selection.

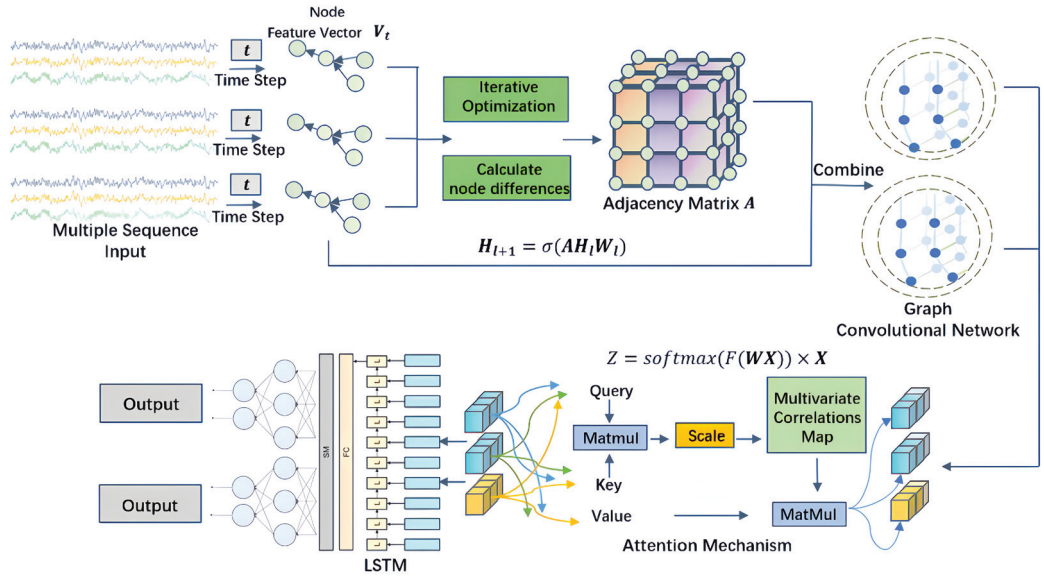


Figure 1. GL-STN schematic diagram.

3.2. Model Algorithm

The algorithm of the GL-STN model is shown in Algorithm 1. Firstly, we initialize the primal variable w and the dual variable v, and set the initial number of nodes N. Subsequently, the algorithm generates a vectorized representation of matrix H by calculating the Euclidean distance between each row of the feature matrix. This step captures the spatial relationships of the data, which is crucial for analyzing the spatiotemporal characteristics of stock market data. Following this, the primal variable w and the dual variable v are iteratively updated to ensure the model can effectively adapt to the dynamic changes in the

market. Finally, the output from the GCN is processed through an attention mechanism. This allows the model to focus on key features to enhance the accuracy and efficiency of predictions.

Algorithm 1. GL-STN (PDS unrolling)

Input: Feature matrix $\boldsymbol{V} \in R^{N \times F}$, number of unrolling times L, input dimension $input_dim$

 # Initialize primal variable $\boldsymbol{w}$, dual variable $\boldsymbol{v}$, and set the initial number of nodes N based on the shape of $\boldsymbol{V}$

Initial primal $\boldsymbol{w_0} = 0 \in R^{N(N-1)/2}$, dual $\boldsymbol{v_0} = 0 \in R^N$ and number of initialization nodes $N = \boldsymbol{V} \cdot shape[0]$

#Calculate the vectorized matrix $\boldsymbol{H}$, where H_{ij} is the squared Euclidean distance between the i and j rows of $\boldsymbol{V}$.

$\boldsymbol{h} = vech(\boldsymbol{H})$ where $H_{ij} = \|\boldsymbol{V_{i,:}} - \boldsymbol{V_{j,:}}\|_2^2$

 # Vectorize and reshape primal variable $\boldsymbol{w}$ into matrix $\boldsymbol{W}$

$\boldsymbol{W} = vech(\boldsymbol{w}, N)$

Degree operator $\boldsymbol{D}$ such that $\boldsymbol{D_W} = [d_1, \ldots, d_N]^T$,where $d_i = \sum_j W_{ij}$ and $\boldsymbol{w} = vech(\boldsymbol{W})$

for $l = 0, 1, \ldots, L$ **do**

$\boldsymbol{r_{1,l}} = \boldsymbol{w_l} - \gamma_l(2\beta_l \boldsymbol{w_l} + 2\boldsymbol{h})$

 # Using degree operator $\boldsymbol{D}$ to update dual variable $\boldsymbol{v}$

$\boldsymbol{r_{2,l}} = \boldsymbol{v_l} + \gamma_l \boldsymbol{D_{wl}}^T \boldsymbol{D_{wl}}$

$\boldsymbol{p_{1,l}} = max\{0, \boldsymbol{r_{1,l}}\}$

 # Using regularization parameters α_l to calculate intermediate variables for dual updates

$\boldsymbol{p_{2,l}} = (\boldsymbol{r_{2,l}} - \sqrt{\boldsymbol{r_{2,l}^2} + 4\alpha_l\gamma_l})/2$

$\boldsymbol{q_{1,l}} = \boldsymbol{p_{1,l}} - \gamma_l\left(2\beta_l \boldsymbol{p_{1,l}} + 2\boldsymbol{h}\right)$

$\boldsymbol{q_{2,l}} = \boldsymbol{p_{2,l}} + \gamma_l \boldsymbol{D_{wl}}^T \boldsymbol{D_{wl}}$

$\boldsymbol{w_{l+1}} = \boldsymbol{w_l} - \boldsymbol{r_{1,l}} + \boldsymbol{q_{1,l}}$

$\boldsymbol{v_{l+1}} = \boldsymbol{v_l} - \boldsymbol{r_{2,l}} + \boldsymbol{q_{2,l}}$

end for

 # Reshape the final primal variable $\boldsymbol{w}$ to the complete adjacency matrix $\boldsymbol{A}$

$\boldsymbol{A_{full}} = vech(\boldsymbol{w}, N)$

$\boldsymbol{A} = \boldsymbol{A_{full}}[:, : input_dim]$

$\boldsymbol{gcn_{input}} = \boldsymbol{A} \times \boldsymbol{V_t}$

$\boldsymbol{gcn_{output}} = \boldsymbol{W} \times \boldsymbol{gcn_{input}}$

$\boldsymbol{attention_{weighs}} = softmax\left(\boldsymbol{W} \times \boldsymbol{gcn_{output}}\right)$

 # Return the attention output

return $attention_{output} = sum\left(\boldsymbol{gcn_{output}} \times \boldsymbol{attention_{weighs}}, axis = 1\right)$

3.3. Multi-Sequence Input

The model framework introduces a multi-sequence input to process the multidimensional characteristics of the stock market. Each sequence represents specific attributes of stock data, such as the price, trading volume, market indicators, etc. Sequences are preprocessed and transformed into node feature vectors V_t at time step t. The multi-dimensional sequence input strategy enables the GL-STN model to comprehensively capture various aspects of the stock market. It can provide a detailed data foundation for the graph structure analysis phase.

3.4. Graph Learning

To reveal the latent connections among stock data, we refer to the Learning to Graph (L2G) theory [28]. Node feature vectors V_t is used to learn the graph adjacency matrix A_t [29,30]. A_t reveals the latent connections between stock market assets. Subsequently,

the adjacency matrix is normalized to produce $\widetilde{A}_t$, as shown in Equation (1), where D_t is the degree matrix derived by summing each row of A_t.

$$\widetilde{\boldsymbol{A}}_{\boldsymbol{t}} = \boldsymbol{D_t}^{-\frac{1}{2}} \boldsymbol{A_t} \boldsymbol{D_t}^{-\frac{1}{2}} \tag{1}$$

After constructing the normalized network, the network features $\widetilde{A}_t$ and asset feature matrix U_t are combined. Asset trends are estimated through a linear combination layer with parameters θ. θ and b are weights learned during training. The output trend estimation y_t is shown in Equation (2).

$$\boldsymbol{y_t} = \widetilde{\boldsymbol{A}}_{\boldsymbol{t}} \boldsymbol{U_t} \boldsymbol{\theta} + \boldsymbol{b} \tag{2}$$

Finally, the sign function sgn is applied to y_t to produce trading signals x_t, as shown in Equation (3). This step transforms trend estimates into trading decisions, building a network-based trading strategy.

$$\boldsymbol{x_t} = sgn(\boldsymbol{y_t}) \tag{3}$$

The application of graph learning enhances the accuracy of stock data analysis, addressing the inadequacies in considering the interconnectivity among individual stocks [31].

3.5. Graph Convolution

In the GL-STN model, GCN is used to analyze complex interactions in stock market data [32,33]. The model's core objective is to extract key features from graph data to deeply understand the influences between nodes. Initially, the adjacency matrix A is calculated based on historical data between stocks. Then, the model performs graph convolution operations by combining the node feature vectors V_t at each time step t with the adjacency matrix A, as shown in Equation (4). H_l represents the node feature matrix of the lth layer, while W_l denotes the corresponding weight matrix. The activation function σ is used to help the model learn complex feature representations.

$$\boldsymbol{H_{l+1}} = \sigma(\boldsymbol{A H_l W_l}) \tag{4}$$

The graph convolution operation can be simplified, as shown in Equation (5), because we use linear activation functions in the model. The feature representation of a node is determined solely by the weighted sum of its neighbors' features, without introducing additional nonlinear transformations.

$$\boldsymbol{H_{l+1}} = \boldsymbol{A W_l} \tag{5}$$

Through GCN operations, the GL-STN model generates updated feature representations for each node, reflecting the stock's relative position and potential market influence in the network [34]. It is crucial for understanding the network role value of each stock.

3.6. Attention Mechanism

The GL-STN model incorporates an attention mechanism to enhance the model's ability to capture and express important information [35,36]. Firstly, linear layers are applied to process the output of GCN. Then, attention coefficients are calculated to generate attention weights for each node, as shown in Equation (6). X represents the node features output by the graph convolutional network, W is the weight matrix of the linear layer, F is a linear transformation function. By using the $softmax$ function, the output of the linear layer is normalized into a probability distribution.

$$AttentionCoefficients = softmax(F(\boldsymbol{WX})) \tag{6}$$

Subsequently, the GL-STN model aggregates the node features weighted by the calculated attention coefficients, as seen in Equation (7). The $\times$ symbol represents element-wise multiplication, and Z is the weighted node feature representation.

$$\boldsymbol{Z} = softmax(F(\boldsymbol{WX})) \times \boldsymbol{X} \tag{7}$$

Through the attention mechanism, the GL-STN model can accurately identify and emphasize key node features, thereby precisely capturing individual stock characteristics and enhancing the accuracy of the feature aggregation phase.

3.7. Temporal Processing

In the GL-STN model, the GCN output weighted by the attention mechanism is fed into the LSTM [37] layer. This enables the model to combine the complex relationships between nodes revealed by the GCN. It also accounts for the dynamic changes in the temporal dimension. LSTM processes inputs over time steps with its recursive structure, retaining memory of past information, thereby extracting richer and deeper features. Subsequently, the output from the LSTM layer is passed to a fully connected layer to form the final prediction result.

The core of the LSTM lies in its cell state and three gate mechanisms (forget gate, input gate, and output gate), which together manage the storage and removal of information. The forget gate determines the information to be retained, the input gate updates the cell state, and the output gate controls the output of the hidden state. Specifically, the gate structure of the LSTM is defined as shown in Equations (8)–(12), where i_t, f_t, c_t, o_t, and h_t represent the input gate, forget gate, memory cell, output gate, and hidden state:

$$i_t = \delta\Big(\boldsymbol{W_i x_t} + \boldsymbol{W_{hi} h_{t-1}} + \boldsymbol{W_{ci} c_{t-1}} + \hat{b}_i\Big) \tag{8}$$

$$f_t = \delta\Big(\boldsymbol{W_f x_t} + \boldsymbol{W_{hf} h_{t-1}} + \boldsymbol{W_{cf} c_{t-1}} + \hat{b}_f\Big) \tag{9}$$

$$i_t = f_t \boldsymbol{c_{t-1}} + i_t tanh\Big(\boldsymbol{W_c x_t} + \boldsymbol{W_{hc} h_{t-1}} + \hat{b}_c\Big) \tag{10}$$

$$o_t = \delta\Big(\boldsymbol{W_o x_t} + \boldsymbol{W_{ho} h_{t-1}} + \boldsymbol{W_{co} c_t} + \hat{b}_o\Big) \tag{11}$$

$$\boldsymbol{h_t} = o_t tanh(\boldsymbol{c_t}) \tag{12}$$

By incorporating LSTM, the GL-STN model significantly improves in recognizing and analyzing the temporal dependencies of data, achieving comprehensive feature learning of stock data in both the temporal and spatial dimensions. This provides strong support for quantitative stock selection.

4. Experimental Evaluation

4.1. Experimental Data

The data we used were sourced from the Tushare platform, covering stocks listed on the Main Board, SME Board, STAR Market, and ChiNext Market from 1 June 2023 to 10 December 2023. Specifically, our dataset encompassed daily trading records of a diverse array of stocks from these different markets, ensuring a comprehensive and representative sample. The Main Board and SME Board collectively include 3201 stocks, generating an aggregate daily trading data volume of 323,721 records. The STAR Market comprises 566 stocks, adding an additional 52,423 daily trading records. The ChiNext Market includes 1332 stocks, contributing a further 152,769 daily trading records. The extensive collection captures the full breadth of market dynamics, providing a robust and solid foundation for our model. For effective training and evaluation, the dataset was divided into a training set (1 June 2023, to 1 August 2023), a validation set (2 August 2023, to 10 August 2023), and a test set (11 August 2023, to 10 December 2023). The model inputs included trading information and derived technical indicators, as detailed in Table 1. By analyzing the input

features, the model outputted probability values indicating the likelihood of stock price increases. This approach offered precise estimations of stock market dynamics.

Table 1. Stock data input by the model.

Data Name	Data Content	Data Name	Data Content
Stock Basic Information	TS Code—ts_code Stock Code—symbol Name—name Listing Date—list_date Delisting Date—delist_date Market Type—market Exchange—exchange HS Mark—is_hs Region—area Industry—industry Full Name—fullname	Index Historical Information	SSE Composite Index—000001.SH SSE 50—000016.SH SSE A Share Index—000002.SH SZSE Component Index—399001.SZ SZSE 300—399007.SZ SME 300—399008.SZ ChiNext 300—399012.SZ SME Board Composite—399101.SZ ChiNext Composite—399102.SZ
Stock Daily Line Information	Stock Code—ts_code Opening Price—open Closing Price—close Highest Price—high Lowest Price—low Price Change—change Percentage Change—pct_chg Volume (shares)—vol Trade Date—trade_date Previous Close—pre_close Transaction Amount—amount	Daily Statistics	Number of Limit Ups—Limit up Number of Limit Downs—Limit down
		Major Trading Markets	SSE Index Information—SSE SZSE Index Information—SZSE
		Individual Stock Money Flow	Flow (money flow)—moneyflow

4.2. Experimental Setup

4.2.1. Experimental Platform

The GL-STN model we propose is implemented using Python 3.5 and the PaddlePaddle framework. All reference models are also implemented using the same tools. The hardware environment for the experiments includes a system with a Tesla V100 GPU, a quad-core CPU, 32 GB of video memory, 32 GB of RAM, and 100 GB of disk storage space. The model training period is set to 1000 epochs, with a fixed learning rate of 0.00001, and each batch contains 190 samples. The loss function for all models involving neural networks is BCEWithLogitsLoss [38], and the optimizer used is AdamW [39].

4.2.2. Baseline Model Comparison

Considering the structural characteristics of the current Chinese stock market, comparative experiments were conducted on the Main Board and SME Board, STAR Market, and ChiNext Market.

The Main Market is a merger of the original Main Board and the SME Board, mainly traded on the Shenzhen Stock Exchange and the Shanghai Stock Exchange. It covers both large mature enterprises and small and medium-sized growth enterprises, representing the core and diversity of the Chinese stock market. The ChiNext Market, located at the Shenzhen Stock Exchange, focuses on supporting enterprises with strong growth potential and innovation capabilities. The STAR Market, located at the Shanghai Stock Exchange, is dedicated to supporting enterprises with core technologies and high innovation capabilities. These three segments collectively promote the diversified development of the Chinese stock market, playing a crucial role in market stability and innovation.

To comprehensively evaluate the effectiveness of the GL-STN model, we selected three representative gradient boosting models (LightGBM [40], CatBoost [41], XGBoost [42]) and three neural network models (LSTM [43], CNN-LSTM [44–48], Transformer [49,50]) for result comparison. Additionally, we conducted ablation experiments to compare the performance differences between three models: a simplified model with only the L2G layer (L2GMOM), an intermediate model combining the L2G layer and LSTM (L2GMOM_LSTM), and the full GL-STN model.

4.2.3. Evaluation Metrics

Evaluation metrics included periodic return, excess return, winning rate, maximum drawdown, Sharpe ratio, Sortino ratio, information ratio, alpha, and beta, defined as per Equations (13)–(21).

p_{t-1} represents the initial capital, p_t denotes the return, R_p is the strategy's annualized return rate, R_m is the benchmark's annualized return rate, and R_f refers to the risk-free rate. Additionally, σ indicates the standard deviation of the portfolio's annualized return rate, D_i is the net value of the product on day i, and D_j is the net value of the product on some day after D_i. D_R represents the downside standard deviation, α is the portfolio's excess return, ω is the active risk, D_p is the strategy's daily return, and D_m represents the benchmark's daily return. NWT is the number of winning trades and TNT is the total number of trades.

$$Periodic\ Return = \frac{p_t - p_{t-1}}{p_{t-1}} \times 100\% \tag{13}$$

$$Excess\ Return = \Big(R_P - R_f\Big) \times 100\% \tag{14}$$

$$Winning\ Rate = \frac{NWT}{TNT} \times 100\% \tag{15}$$

$$Maximum\ Drawtown = max\left(\frac{D_i - D_j}{D_i}\right) \tag{16}$$

$$Sharpe\ Ratio = \frac{R_p - R_f}{\sigma} \times 100\% \tag{17}$$

$$Sortino\ Ratio = \frac{R_p - R_f}{D_R} \tag{18}$$

$$Information\ Ratio = \frac{\alpha}{\omega} \tag{19}$$

$$Alpha = R_p - \Big[R_f + \beta \times \Big(R_m - R_f\Big)\Big] \tag{20}$$

$$Beta = \frac{Cov(D_p, D_m)}{Var(D_m)} \tag{21}$$

4.3. Experimental Results and Analysis

4.3.1. Main Board and SME Board Experimental Results Comparison

The Main Board and SME Board markets are known for their mature trading environment and high volatility. Among the gradient boosting models, LightGBM leads in periodic return, with a performance of -0.0930, while CatBoost and XGBoost have return values of -0.1018 and -0.1484, respectively. However, the average winning rate of these models is only 0.3921, highlighting their limitations in providing effective trading decision support. Risk-adjusted return metrics, including average Sharpe ratio, Sortino ratio, and information ratio, are -2.1989, -3.5394, and -0.7376. These results further highlight the limitations in risk-adjusted returns and information generation, as shown in Table 2.

In neural network models, LSTM, CNN-LSTM, and Transformer show instability in maintaining account value with periodic return metrics of -0.0759, -0.1379, and -0.1309. The winning rate metrics highlight shortcomings in trading decision-making—0.3913 for

LSTM, 0.3194 for CNN-LSTM, and 0.3846 for Transformer. The average information ratio for these three models is about −0.9868, suggesting deficiencies in providing efficient trading decisions and decision-supporting information, as shown in Table 2.

Table 2. The performance of different models on various metrics based on Main Board and SME Board.

Metric	GL-STN	LSTM	CNN-LSTM	Transformer	CatBoost	LightGBM	XGBoost
Periodic Return	0.0343	−0.0759	−0.1379	−0.1309	−0.1018	−0.0930	−0.1484
Excess Return	1.000	−0.0102	−0.0722	−0.0652	−0.0361	−0.0273	−0.0827
Winning Rate	0.5581	0.3913	0.3194	0.3846	0.3509	0.4274	0.3982
Maximum Drawdown	0.0221	0.0898	0.1379	0.1463	0.1384	0.1030	0.1541
Sharpe Ratio	1.1929	−3.3520	−4.7526	−3.2786	−1.8131	−1.7436	−3.0400
Sortino Ratio	2.1403	−4.6086	−6.4848	−5.0680	−3.4037	−2.8529	−4.3617
Information Ratio	2.6910	−0.2072	−1.5660	−1.1874	−0.4921	−0.3558	−1.3650
Alpha	0.1314	−0.2392	−0.4940	−0.4597	−0.2917	−0.2851	−0.5053
Beta	0.0941	0.0428	−0.1185	−0.1001	0.1479	0.0453	−0.0341

In contrast, the GL-STN model outperforms other models with a positive periodic return metric of 0.0343, demonstrating its performance in maintaining account value stability. Additionally, GL-STN achieves a winning rate of 0.5581, a maximum drawdown of only 0.0221, and significantly high values for the Sharpe ratio and Sortino ratio. The account value change curves of different models, as depicted in Figure 2, visually underscore the GL-STN model's superior ability to sustain account value in the volatile markets of the Main Board and SME Board. Additionally, the ablation experiments, illustrated in Figure 3, demonstrate the efficiency and robustness of the GL-STN model in navigating market complexities. This is shown through the comparative account value change curves of the models under stringent tests. These experiments, as detailed in Tables 3 and 4, affirm the GL-STN model's standout performance and high efficiency.

Moreover, Figure 4 provides a radar comparison chart, which succinctly contrasts the comprehensive metrics of various models, highlighting GL-STN's outstanding performance across multiple dimensions, thereby reinforcing its superiority in navigating the complexities of high-volatility trading environments.

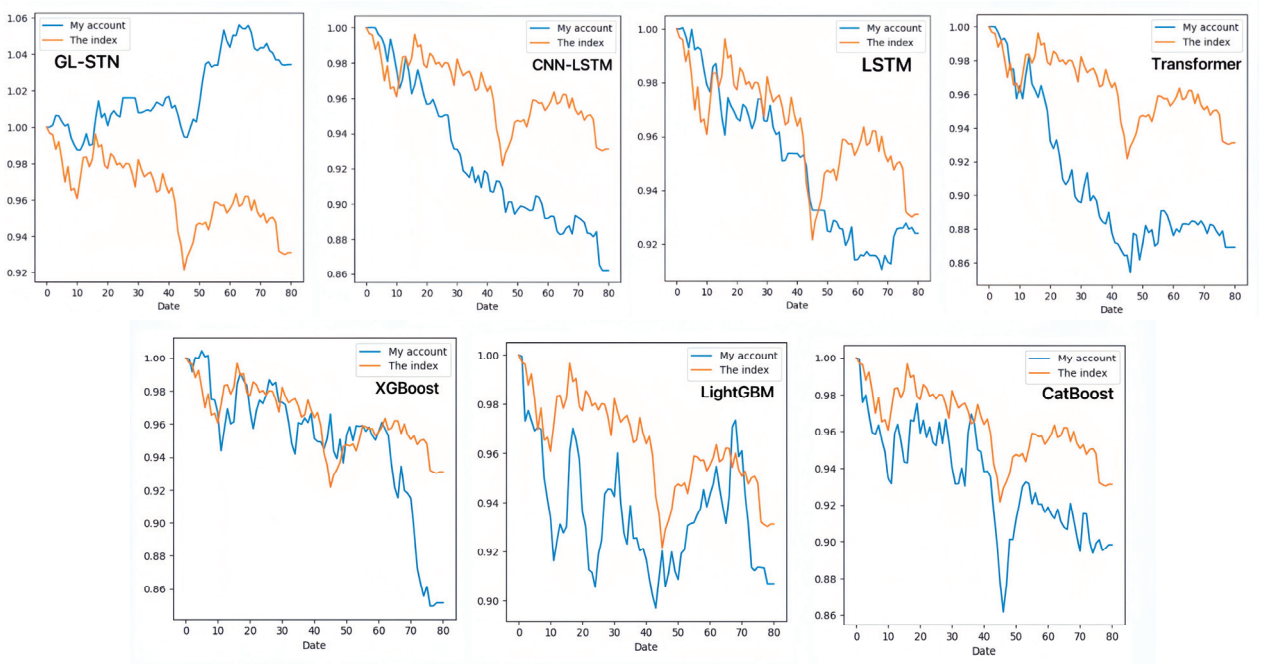


Figure 2. The account value change curves of different models based on Main Board and SME Board.

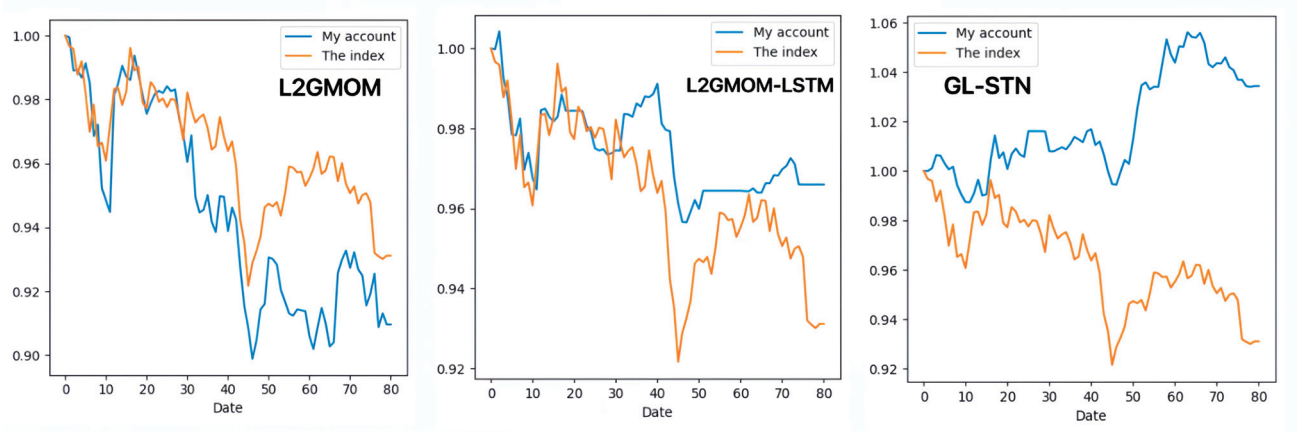


Figure 3. The account value change curves of models in the ablation experiment based on Main Board and SME Board.

Table 3. The performance of models in the ablation experiment on various metrics based on Main Board and SME Board.

Metric	L2GMOM	L2GMOM_LSTM	GL-STN
Periodic Return	−0.0904	−0.0341	0.0343
Excess Return	−0.0247	0.0316	1.000
Winning Rate	0.4059	0.3667	0.5581
Maximum Drawdown	0.1012	0.0475	0.0221
Sharpe Ratio	−2.1648	−1.7905	1.1929
Sortino Ratio	−3.6846	−2.1141	2.1403
Information Ratio	−0.4155	1.0410	2.6910
Alpha	−0.2606	−0.0639	0.1314
Beta	0.1421	0.1992	0.0941

Table 4. Runtime of models in the ablation experiment based on Main Board and SME Board.

Model	Runtime/h
L2GMOM	45.1663
L2GMOM_LSTM	41.2852
GL-STN	27.5765

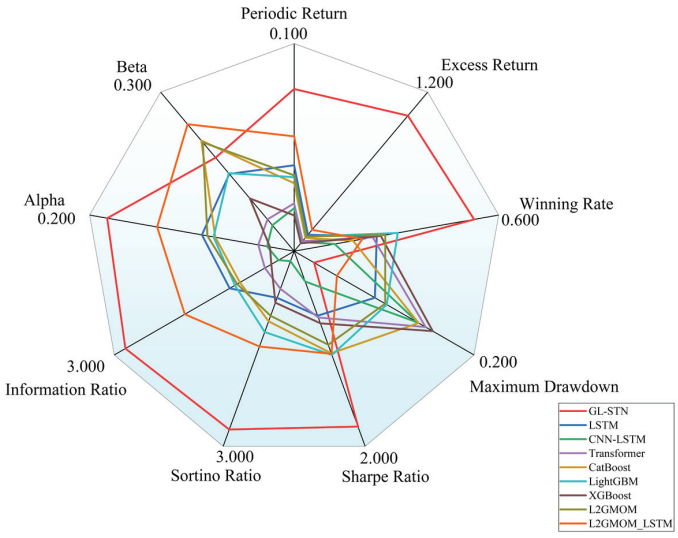


Figure 4. Radar comparison chart of model metrics on Main Board and SME Board.

4.3.2. STAR Market Experimental Results Comparison

In the science and technology innovation-driven STAR Market, gradient boosting models (LightGBM, CatBoost, XGBoost) show limitations in decision-making quality amidst high volatility. All models exhibit negative periodic returns and an average trade win rate of approximately 0.3837. The Sharpe ratio, Sortino ratio, and information ratio averages of −2.6174, −4.0004, and −0.7846, respectively, underscore the challenges in achieving risk-adjusted returns and generating informative insights, as shown in Table 5.

Table 5. The performance of different models on various metrics based on STAR Market.

Metric	GL-STN	LSTM	CNN-LSTM	Transformer	CatBoost	LightGBM	XGBoost
Periodic Return	0.0056	−0.0808	−0.1001	−0.0861	−0.1232	−0.0972	−0.1085
Excess Return	0.0713	−0.0151	−0.0343	−0.0204	−0.0575	−0.0315	−0.0428
Winning Rate	0.4615	0.1852	0.3929	0.2174	0.3803	0.4215	0.3494
Maximum Drawdown	0.0162	0.0808	0.1185	0.0861	0.1604	0.1160	0.1235
Sharpe Ratio	0.0014	−4.4737	−2.0821	−5.9939	−2.9655	−1.9384	−2.9483
Sortino Ratio	0.0016	−5.0342	−3.4462	−6.4179	−4.7291	−2.8752	−4.3969
Information Ratio	1.9907	−0.3680	−0.5245	−0.5275	−1.0092	−0.4724	−0.8724
Alpha	0.0066	−0.2593	−0.3128	−0.2919	−0.4305	−0.2728	−0.3265
Beta	−0.0551	0.0330	0.0452	−0.0273	−0.0977	0.1751	0.1422

Neural network models encounter similar hurdles in the STAR Market. The LSTM and CNN-LSTM models report periodic returns of −0.0808 and −0.1001, respectively, while the Transformer model slightly outperforms them at −0.0861. This performance indicates instability in maintaining account value, compounded by a low average trade winning rate of 0.2651 and an average information ratio of −0.4733. These results, shown in Table 5, highlight deficiencies in delivering precise trading strategies and supporting information.

Contrastingly, the GL-STN model demonstrates exceptional stability in the STAR Market, reflected by a positive periodic return of 0.0056, a winning rate of 0.4615, and a maximum drawdown rate of 0.0162, alongside impressive Sharpe and Sortino ratios. Its performance distinctly surpasses that of other models, underlining its efficiency in decision-making, account value stability, and risk management. The account value change curves for different models on the STAR Market, illustrated in Figures 5 and 6, visually emphasize the GL-STN model's resilience and superior ability to maintain account value against its counterparts. Furthermore, Tables 6 and 7 underscore the robustness of the GL-STN model in ablation studies conducted on the STAR Market. They also demonstrate its operational efficiency, with a runtime of only 4.7232 h. Figure 7's radar comparison chart of model metrics on the STAR Market offers a comprehensive visualization of the GL-STN model's outstanding performance across various dimensions, affirming its advanced capability in managing the intricacies of high-volatility trading environments.

Table 6. The performance of models in the ablation experiment on various metrics based on STAR Market.

Metric	L2GMOM	L2GMOM_LSTM	GL-STN
Periodic Return	−0.0991	−0.0261	0.0056
Excess Return	−0.0334	0.0396	0.0713
Winning Rate	0.3838	0.3913	0.4615
Maximum Drawdown	0.1300	0.0543	0.0162
Sharpe Ratio	−2.1716	−0.9198	0.0014
Sortino Ratio	−3.1699	−1.3231	0.0016
Information Ratio	−0.5372	0.9734	1.9907
Alpha	−0.3041	−0.0679	0.0066
Beta	0.0753	0.0495	−0.0551

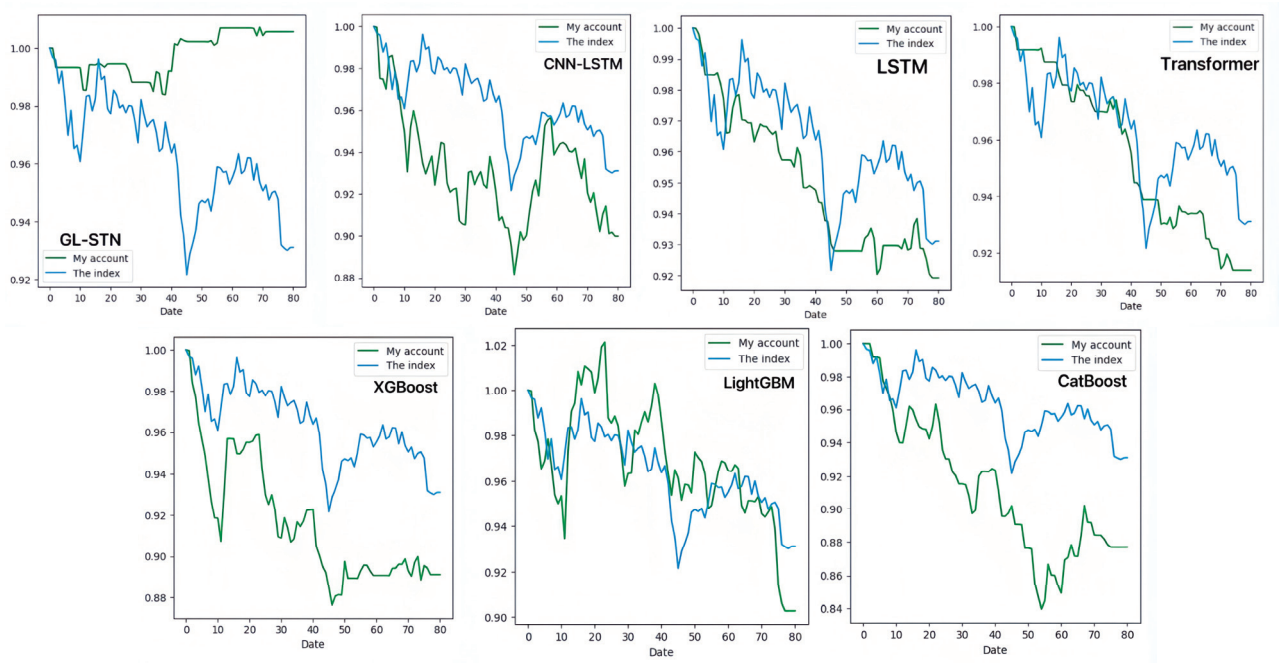


Figure 5. The account value change curves of different models based on STAR Market.

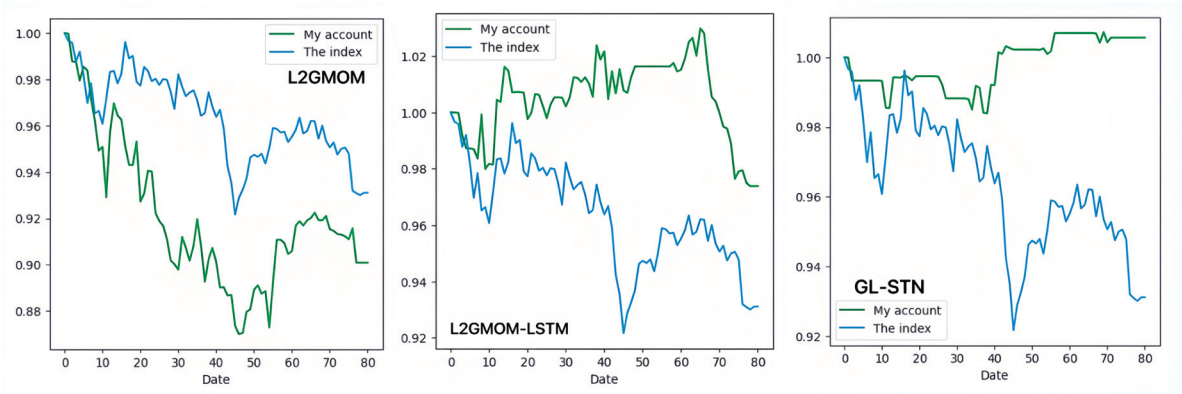


Figure 6. The account value change curves of models in the ablation experiment based on STAR Market.

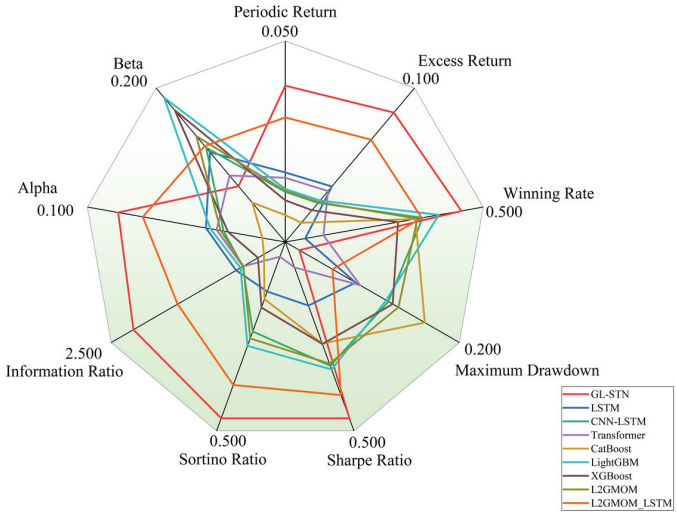


Figure 7. Radar comparison chart of model metrics on STAR Market.

Table 7. Runtime differences in models in the ablation experiment based on STAR Market.

Model	Runtime/h
L2GMOM	6.9918
L2GMOM_LSTM	6.9294
GL-STN	4.7232

4.3.3. ChiNext Market Experimental Results Comparison

In the ChiNext Market, gradient boosting models have shown consistent performance in periodic return but overall poor effectiveness. LightGBM slightly outperforms the others with a result of -0.0861, while CatBoost and XGBoost achieve values of -0.0966 and -0.0744. Moreover, the average values for the Sharpe ratio, Sortino ratio, and Information ratio are -1.8532, -2.7696, and -0.2534, indicating deficiencies in profitability and information output after adjusting for risk, as shown in Table 8.

Table 8. The performance of different models on various metrics based on ChiNext Market.

Metric	GL-STN	LSTM	CNN-LSTM	Transformer	CatBoost	LightGBM	XGBoost
Periodic Return	−0.0067	−0.0917	−0.0742	−0.0778	−0.0966	−0.0861	−0.0744
Excess Return	0.0590	−0.0260	−0.0085	−0.0121	−0.0309	−0.0204	−0.0087
Winning Rate	0.4722	0.4333	0.4342	0.4000	0.3929	0.3824	0.4103
Maximum Drawdown	0.0728	0.1011	0.0942	0.1374	0.1268	0.1068	0.1115
Sharpe Ratio	−0.3553	−2.0296	−1.9864	−1.8095	−1.9718	−2.2814	−1.3065
Sortino Ratio	−0.5154	−2.6046	−3.0227	−3.3699	−3.1082	−3.1015	−2.0993
Information Ratio	1.5242	−0.4062	−0.0933	−0.1441	−0.4675	−0.3258	0.0330
Alpha	0.0138	−0.2686	−0.2499	−0.2377	−0.2785	−0.2986	−0.2239
Beta	0.1366	0.1185	−0.0546	0.0443	0.1427	−0.0903	0.0226

In the evaluation of neural network models, LSTM and CNN-LSTM failed to demonstrate stability in periodic return, while the Transformer model showed slight improvement with a fluctuation value of -0.0778. Additionally, the average trade winning rate for these models is only 0.4225, suggesting limited capability in providing efficient trading decision support. The average information ratio of -0.2145 further reveals the models' inadequacy in delivering effective decision-support information, as shown in Table 8.

Conversely, the GL-STN model emerges as a standout in the ChiNext Market, showcasing significant stability and superior performance. It leads with a winning rate of 0.4722, underscoring its precision in trading decisions. It also exemplifies unparalleled loss control capabilities with a maximum drawdown rate of 0.0728. Furthermore, GL-STN's superior Sharpe and Sortino ratios reflect its efficacy in securing higher returns at minimized risks. With an information ratio of 1.5242, GL-STN stands at the forefront, demonstrating its exceptional ability to outperform market benchmarks. The model's excellence extends to the alpha and beta indicators, affirming its superior risk-adjusted returns and market adaptability.

The account value change curves for different models in the ChiNext Market, shown in Figures 8 and 9, clearly highlight the GL-STN model's robustness and consistent account value preservation, setting it apart from its peers. The ablation experiment results, as presented in Tables 9 and 10, demonstrate that GL-STN outperforms both L2GMOM and L2GMOM_LSTM models in terms of its shorter running time and superior index benefits. This further substantiates its computational efficiency and exceptional performance. The comprehensive radar comparison chart in Figure 10 encapsulates the model metrics on the ChiNext Market, offering a succinct overview of GL-STN's leading performance across various dimensions.

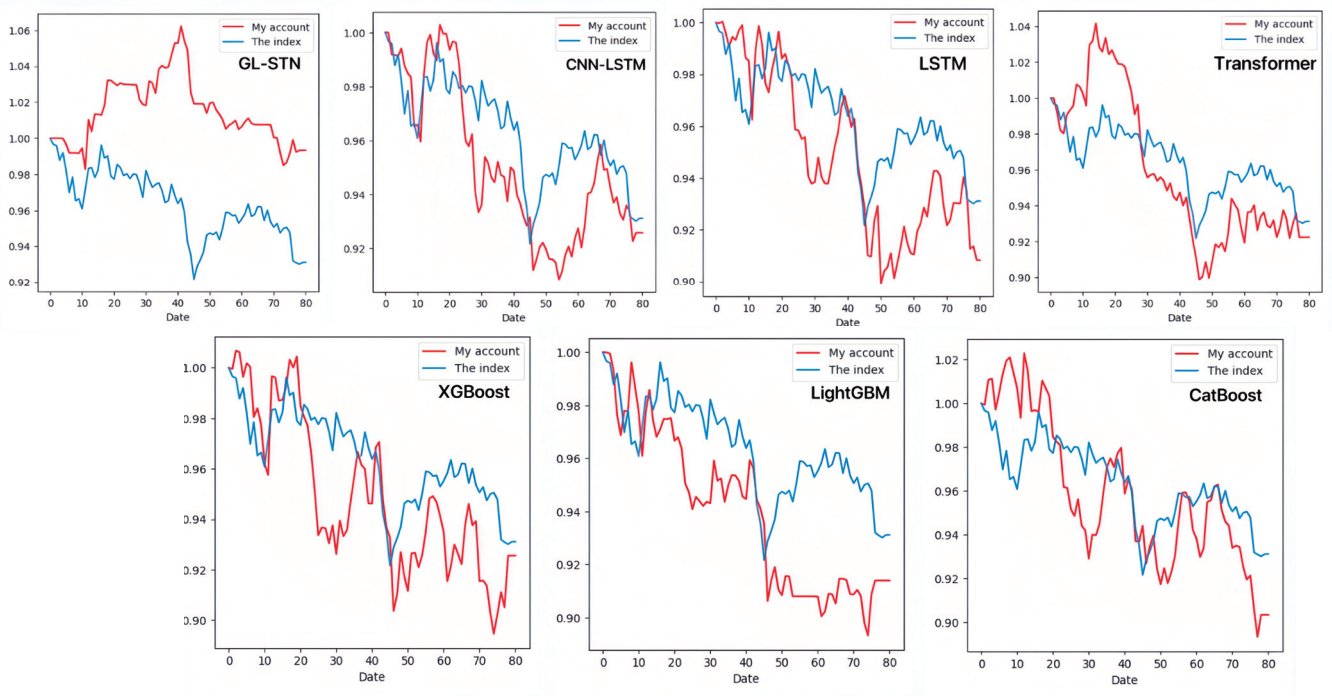


Figure 8. The account value change curves of different models based on ChiNext Market.

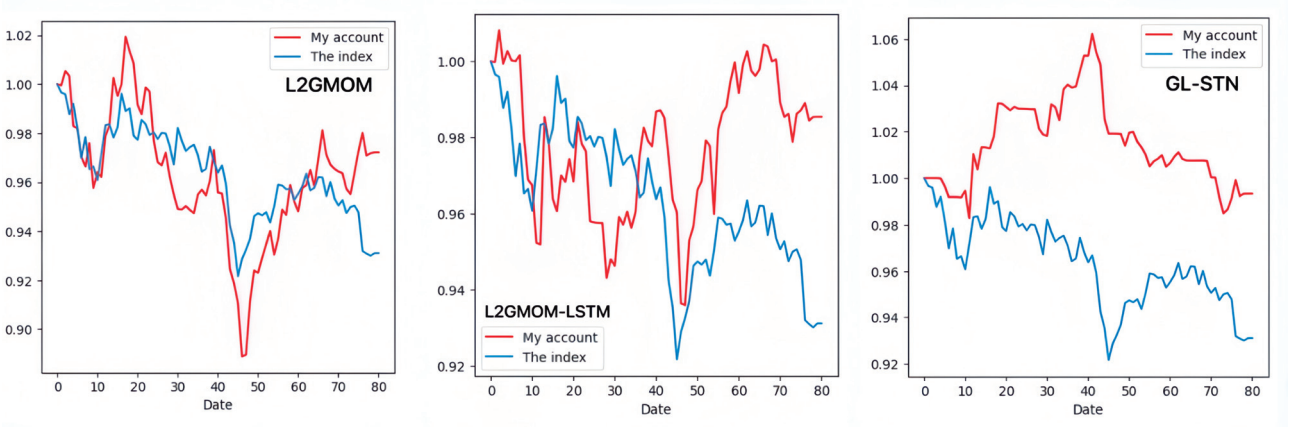


Figure 9. The account value change curves of models in the ablation experiment based on ChiNext Market.

Table 9. The performance of models in the ablation experiment on various metrics based on ChiNext Market.

Metric	L2GMOM	L2GMOM_LSTM	GL-STN
Periodic Return	−0.0279	−0.0146	−0.0067
Excess Return	0.0379	0.0511	0.0590
Winning Rate	0.4712	0.4756	0.4722
Maximum Drawdown	0.1281	0.0717	0.0728
Sharpe Ratio	−0.6081	−0.3555	−0.3553
Sortino Ratio	−0.9346	−0.5264	−0.5154
Information Ratio	0.7931	1.0305	1.5242
Alpha	−0.0452	−0.0181	0.0138
Beta	0.1467	0.0783	0.1366

Table 10. Runtime differences in models in the ablation experiment based on ChiNext Market.

Model	Runtime/h
L2GMOM	34.0752
L2GMOM_LSTM	33.9639
GL-STN	24.2041

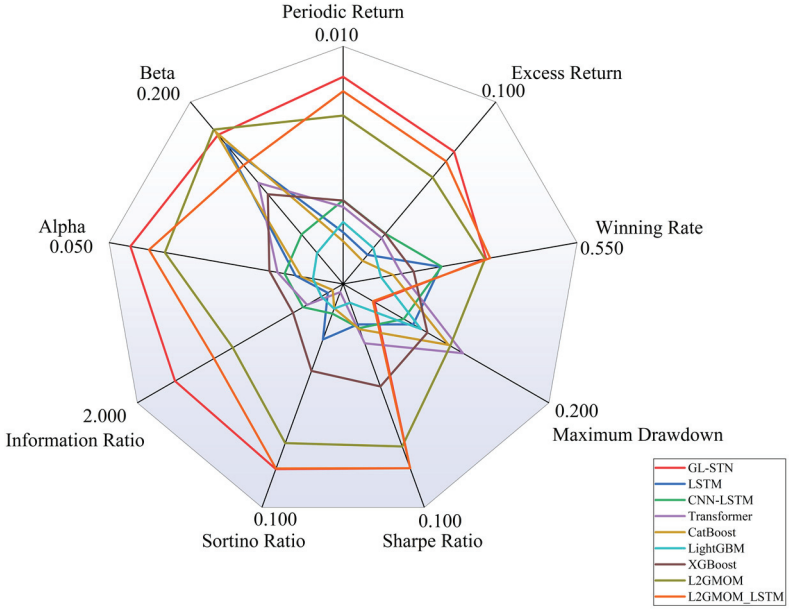


Figure 10. Radar comparison chart of model metrics on ChiNext Market.

5. Discussion and Analysis

Building on the evaluation of the GL-STN model's performance across different Chinese stock market segments, it is crucial to emphasize the Chinese stock market's representativeness in the global financial landscape. As one of the world's largest stock markets, the Chinese stock market is representative not only in terms of scale and liquidity but also in volatility, regulatory environment, market structure, and technological applications. It offers a comprehensive landscape for evaluating the GL-STN model and provides valuable insights into global financial markets.

Firstly, China's Main Board market encompasses a large number of mature, large-cap companies. It has relatively low volatility and high liquidity, similar to the New York Stock Exchange (NYSE) and National Association of Securities Dealers Automated Quotations (NASDAQ) in the United States. The SME board mainly represents the investment environment for small and medium-sized enterprises, akin to the Frankfurt Stock Exchange (FSE) and other European stock markets. These markets offer an ideal environment for testing the model's adaptability to different economic scales and types of enterprises. On the Main and SME Boards, the GL-STN model demonstrates effective market volatility control and strong profitability through lower cyclical returns and higher Sharpe ratios. These characteristics indicate that the GL-STN model has the potential to perform well in other mature markets with similar attributes globally.

Secondly, the STAR Market and ChiNext Market are characterized by high volatility and high growth, making them suitable for studying model performance in emerging market environments. Emerging markets such as India and Brazil also share these characteristics, with high innovation potential but also high risk. On the STAR Market, the GL-STN model maintains stable trading success rates and controlled maximum drawdowns. In the ChiNext Market, the GL-STN model shows good adaptability, effectively handling market volatility and risk management. This implies its potential applicability in other high-risk, high-reward markets worldwide. By testing in China's diverse markets,

we can better understand and predict the applicability and performance of the GL-STN model in other global markets.

Furthermore, abrupt market shifts and black swan events are crucial factors that cannot be ignored in quantitative stock selection. The GL-STN model incorporates certain measures in the hope of addressing these challenging issues. Firstly, through multi-sequence inputs, the GL-STN model can dynamically adapt to market changes, enhancing its sensitivity to sudden events. Secondly, GL-STN uses GCN to process complex interactions in stock market data, generating feature representations that reflect the relative positions and influences of stocks within the market network. This captures nonlinear and dynamic relationships, improving the model's responsiveness to sudden changes. Additionally, the integrated attention mechanism significantly enhances the GL-STN model's ability to capture and emphasize critical market information, allowing the model to focus on key market signals and maintain robustness in volatile markets. Finally, by combining LSTM to process weighted GCN outputs, GL-STN has the potential to quickly adjust predictions and decisions, demonstrating high flexibility and adaptability during sudden events or black swan incidents.

Moreover, the GL-STN model's robust and complex design, which includes GCN, L2G mechanisms, attention mechanisms, and LSTM, is well-suited for handling large volumes of market data. This architecture effectively captures and analyzes both spatial and temporal features in financial datasets. The design of the GL-STN model facilitates high-dimensional data processing and robust feature extraction, ensuring precise analysis even with extensive and complex datasets.

Additionally, the model structure is highly extendable. Its architecture is designed to accommodate increasing data complexity and volume, making it capable of managing larger datasets and more intricate financial instruments. The modular approach, incorporating components such as GCN and LSTM, allows for the seamless integration of additional layers or mechanisms to enhance performance. This flexibility allows the GL-STN model to adapt to evolving market conditions, integrate new data sources, and incorporate advanced analytical techniques as they become available. Its extensibility supports continuous improvement and adaptation, ensuring its relevance and effectiveness in the ever-changing financial markets.

Overall, the evaluations suggest that the GL-STN model exhibits a degree of versatility and robustness. It shows potential in quantitative stock selection across diverse global markets, demonstrates the ability to handle large datasets, and may assist in managing abrupt market shifts and black swan events.

6. Conclusions and Outlook

The aim of this study was to explore a novel quantitative stock selection method based on the GL-STN model. We conducted an in-depth analysis of traditional quantitative stock selection methods and the principles of the L2G layer. By integrating GCN, L2G, attention mechanisms, and LSTM, we proposed an improved composite investment strategy. Empirical results demonstrate that in various market segments such as the Main Board, SME Board, STAR Market, and ChiNext Market, the GL-STN model exhibits significant stability, profitability, and excellent market adaptability.

In the mature Main and SME Boards, the GL-STN model effectively controlled market volatility and ensured profitability, with a lower periodic return and higher Sharpe ratio. In the volatile STAR Market, the model maintained a stable winning rate and controlled maximum drawdown. In the ChiNext market, it demonstrated strong adaptability in managing market fluctuations and risk.

Despite positive results, there remains room for improvement in model parameter optimization, feature selection, and time efficiency [51]. Future research could explore more efficient parameter adjustment methods and more refined feature engineering techniques to enhance the model's accuracy and generalizability. Furthermore, although the GL-STN model has reduced runtime, there is still scope for further improvement in time efficiency.

Additionally, while the current model focuses mainly on the Chinese stock market, future studies should apply it to international stock markets to explore the model's performance in different market environments. This would not only validate the model's robustness but also provide references for international quantitative investment strategies.

Author Contributions: T.C., conceptualization, methodology, software, writing—original draft; X.W., validation, software; H.W., data curation, resources; X.Y., project administration, supervision; L.X., conceptualization, methodology, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Philosophy and Social Science Planning Cross-disciplinary Key Support Subjects of Zhejiang Province (No. 22JCXK08Z), Ningbo Natural Science Foundation (No. 2022J162), Zhejiang Province Statistical Research Project (No. 23TJZZ22), Ningbo Philosophy and Social Science Research Base Project (No. JD6-228).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the first author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
2. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
3. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
4. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
5. Zhao, H.; Wu, L.; Li, Z.; Zhang, X.; Liu, Q.; Chen, E. Prediction of Internet Financial Market Dynamics Based on Deep Neural Network Structure. *J. Comput. Res. Dev.* **2019**, *56*, 1621–1631.
6. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 802–810.
7. Bollerslev, T. Generalized Autoregressive Conditional Heteroskedasticity. *J. Econom.* **1986**, *31*, 307–327. [CrossRef]
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
9. Wang, S.; Cao, Z.; Chen, M. Research on the Application of Random Forest in Quantitative Stock Selection. *Oper. Res. Manag. Sci.* **2016**, *25*, 163–168+177.
10. Li, B.; Lin, Y.; Tang, W. ML-TEA: A Set of Quantitative Investment Algorithms Based on Machine Learning and Technical Analysis. *Syst. Eng.—Theory Pract.* **2017**, *37*, 1089–1100.
11. Zhou, L. Research on Multi-factor Stock Selection Strategy Based on Quantile Regression. *J. Southwest Univ. Nat. Sci. Ed.* **2019**, *41*, 89–96.
12. Shu, S.; Li, L. Regularized Sparse Multi-factor Quantitative Stock Selection Strategy. *Comput. Eng. Appl.* **2021**, *57*, 110–117.
13. Lewellen, J. The Cross-Section of Expected Stock Returns. *Crit. Financ. Rev.* **2015**, *4*, 1–44. [CrossRef]
14. DeMiguel, V.; Martin-Utrera, A.; Nogales, F.J. A Transaction-Cost Perspective on the Multitude of Firm Characteristics. *Rev. Financ. Stud.* **2017**, *33*, 2180–2222. [CrossRef]
15. Gu, S.; Kelly, B.; Xiu, D. Empirical Asset Pricing Via Machine Learning. *Rev. Financ. Stud.* **2018**, *33*, 32–47.
16. Yu, L.; Wang, S.; Lai, K.K. Mining Stock Market Tendency Using GA-Based Support Vector Machines. In Proceedings of the First International Workshop on Internet and Network Economics: LNCS Volume 3828, Hong Kong, China, 15–17 December 2005; Springer: Berlin, Germany, 2005; pp. 336–345.
17. Mndawe, S.T.; Paul, B.S.; Doorsamy, W. Development of a Stock Price Prediction Framework for Intelligent Media and Technical Analysis. *Appl. Sci.* **2022**, *12*, 719. [CrossRef]
18. Lin, J.; Gong, Z. A Research on Forecasting of Shanghai Zinc Futures Price Based on Artificial Neural Network. *Theory Pract. Financ. Econ.* **2017**, *38*, 54–57.
19. Liu, S.; Zhang, C.; Ma, J. CNN-LSTM Neural Network Model for Quantitative Strategy Analysis in Stock Markets. In Proceedings of the International Conference on Neural Information Processing, Guangzhou, China, 14–18 November 2017; Springer: Cham, Switzerland, 2017; pp. 199–205.
20. Liang, J.; Huang, K.; Qiu, S.; Lin, H.; Lian, K. Trade Filtering Method for Trend Following Strategy Based on LSTM-Extracted Feature and Machine Learning. *J. Intell. Fuzzy Syst.* **2022**, *44*, 6131–6149. [CrossRef]
21. Kim, S.-H.; Park, D.-Y.; Lee, K.-H. Hybrid Deep Reinforcement Learning for Pairs Trading. *Appl. Sci.* **2022**, *12*, 944. [CrossRef]

22. Staffini, A. Stock Price Forecasting by a Deep Convolutional Generative Adversarial Network. *Front. Artif. Intell.* **2022**, *5*, 837596. [CrossRef] [PubMed]
23. Liu, H.; Peng, L.; Tang, Y. Retail Attention, Institutional Attention. *J. Financ. Quant. Anal.* **2023**, *58*, 1005–1038. [CrossRef]
24. Lai, T.-Y.; Cheng, W.J.; Ding, J.-E. Sequential Graph Attention Learning for Predicting Dynamic Stock Trends. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023.
25. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
26. Wang, J.; Zhang, S.; Xiao, Y.; Song, R. A Review on Graph Neural Network Methods in Financial Applications. *J. Data Sci. JDS* **2022**, *20*, 111–134. [CrossRef]
27. Tan, Z.; Hu, M.; Wang, Y.; Wei, L.; Liu, B. Futures Quantitative Investment with Heterogeneous Continual Graph Neural Network. *arXiv* **2023**, arXiv:2303.16532.
28. Pu, X.; Zohren, S.; Roberts, S.; Dong, X. Learning to Learn Financial Networks for Optimising Momentum Strategies. *arXiv* **2023**, arXiv:2308.12212.
29. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Amodei, D. Language Models Are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
30. Hamilton, W.; Ying, Z.; Leskovec, J. Inductive Representation Learning on Large Graphs. *Adv. Neural Inf. Process. Syst.* **2017**, *31*, 1024–1034.
31. Ni, P.; Okhrati, R.; Guan, S.; Chang, V. Knowledge Graph and Deep Learning-Based Text-to-GraphQL Model for Intelligent Medical Consultation Chatbot. *Inf. Syst. Front.* **2024**, *26*, 137–156. [CrossRef] [PubMed]
32. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
33. Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful Are Graph Neural Networks? In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
34. Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph Attention Networks. In Proceedings of the 6th International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
35. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
36. Luong, M.T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-Based Neural Machine Translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015.
37. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
38. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006.
39. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
40. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 3146–3157.
41. Dorogush, A.V.; Gulin, A.; Gusev, G.; Kazeev, N.; Ostroumova Prokhorenkova, L.; Vorobev, A. Fighting Biases with Dynamic Boosting. In Proceedings of the NeurIPS, Montreal, QC, USA, 3–8 December 2018.
42. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
43. Bao, Y.; Sun, Y. A Deep Learning Framework for Financial Time Series Using Stacked Autoencoders and Long-Short Term Memory. *PLoS ONE* **2017**, *12*, e0180944. [CrossRef] [PubMed]
44. Li, T.; Hua, M.; Wu, X. A Hybrid CNN-LSTM Model for Forecasting Particulate Matter (PM2.5). *IEEE Access* **2020**, *8*, 26933–26940. [CrossRef]
45. Dong, Y.; Yan, D.; Almudaifer, A.I.; Yan, S.; Jiang, Z.; Zhou, Y. BELT: A Pipeline for Stock Price Prediction Using News. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 1137–1146.
46. Lucarelli, G.; Borrotti, M. A Deep Reinforcement Learning Approach for Automated Cryptocurrency Trading. In Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations, Hersonissos, Greece, 24–26 May 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 247–258.
47. Wang, J.; Zhang, Y.; Tang, K.; Wu, J.; Xiong, Z. Alphastock: A Buying-Winners-and-Selling-Losers Investment Strategy Using Interpretable Deep Reinforcement Attention Networks. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 1900–1908.
48. Xiong, Z.; Liu, X.-Y.; Zhong, S.; Yang, H.; Walid, A. Practical Deep Reinforcement Learning Approach for Stock Trading. *arXiv* **2018**, arXiv:1811.07522.
49. Rasheed, J.; Jamil, A.; Hameed, A.A.; Ilyas, M.; Özyavaş, A.; Ajlouni, N. Improving Stock Prediction Accuracy Using CNN and LSTM. In Proceedings of the 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), Sakheer, Bahrain, 26–27 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–5.

50. Li, Y.; Pan, Y. A Novel Ensemble Deep Learning Model for Stock Prediction Based on Stock Prices and News. *Int. J. Data Sci. Anal.* **2022**, *13*, 139–149. [CrossRef] [PubMed]
51. Leung, M.-F.; Wang, J.; Che, H. Cardinality-constrained portfolio selection via two-timescale duplex neurodynamic optimization. *Neural Netw.* **2022**, *153*, 399–410. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

MDPI Books Editorial Office
E-mail: books@mdpi.com
www.mdpi.com/books



Disclaimer/Publisher's Note: The title and front matter of this reprint are at the discretion of the Topic Editors. The publisher is not responsible for their content or any associated concerns. The statements, opinions and data contained in all individual articles are solely those of the individual Editors and contributors and not of MDPI. MDPI disclaims responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.





Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-5184-3