ACE

Applied and Computational Engineering

Proceedings of the 6th International Conference
on Computing and Data Science

Portsmouth, UK

September 12 - September 19, 2024

EWA Publishing

Volume 87

Editors
Alan Wang
University of Auckland

Roman Bauer
University of Surrey

ISSN: 2755-2721
ISSN: 2755-273X (eBook)

ISBN: 978-1-83558-585-6
ISBN: 978-1-83558-586-3 (eBook)

Publication of record for individual papers is online:
https://www.ewadirect.com/proceedings/ace/home/index

This imprint is published by **EWA Publishing**
Address: John Eccles House, Robert Robinson Avenue, Oxford, England, OX4 4GP
Email: info@ewapublishing.org

Committee Members

CONF-CDS 2024

General Chair

Elisavet Andrikopoulou, *University of Portsmouth*

Technical Program Chair

Marwan Omar, *Illinois Institute of Technology*

Technical Program Committee

Anil Fernando, *University of Strathclyde*
Michael Harre, *The University of Sydney*
Lewis Tseng, *Clark University*
Mukhtar Ullah, *FAST NUCES Islamabad*
Sameena Naaz, *Jamia Hamdard*
Kourosh Khoshelham, *University of Melbourne*
Weijia Cao, *Aerospace Information Research Institute, Chinese Academy of Sciences*
Zachary Ziegler, *Harvard University*
Haidong Xie, *China Academy of Space Technology*
Ammar Alazab, *Torrens University Australia*

Organizing Chairs

Alan Wang, *University of Auckland*
Roman Bauer, *University of Surrey*

Organizing Committee

Festus Adedoyin, *Bournemouth University*
Haseeb Ahmad, *National Textile University Faisalabad*
Richa Gupta, *Jamia Hamdard*
Tauqir Nasir, *Eastern Mediterranean University*
Alex Siow, *National University of Singapore*
Guozheng Rao, *Tianjing University*
Ce Li, *China University of Mining and Technology*
Brajendra Panda, *University of Arkansas*
Long Bao, *Apple Inc.*
Guoqing Xiang, *Peking University*
Karen Works, *Florida State University*
Rahul Kumar Dubey, *Robert Bosch GmbH*

Lei Shu, *University of Texas at Austin*
Xinqing Xiao, *China Agricultural University*
Wang Juan, *Beijing Computer Federation*

Publicity Chairs

Shuang-Hua Yang, *University of Reading*
Dadmehr Rahbari, *Tallinn University of Technology*
James Duncan-Brown, *University of South Africa*

Preface

The 6th International Conference on Computing and Data Science (CONF-CDS 2024) is an annual conference focusing on research areas including computing technology, machine learning, computer science, and data science. It aims to establish a broad and interdisciplinary platform for experts, researchers, and students worldwide to present, exchange, and discuss the latest advance and development in computing technology, machine learning, computer science, and data science.

This volume contains the papers of the 6th International Conference on Computing and Data Science (CONF-CDS 2024). Each of these papers has gained a comprehensive review by the editorial team and professional reviewers. Each paper has been examined and evaluated for its theme, structure, method, content, language, and format.

Cooperating with prestigious universities, CONF-CDS 2024 organized five workshops in Portsmouth, Chicago, Melbourne and Beijing. Dr. Elisavet Andrikopoulou chaired the workshop "Data Visualization Methods in Healthcare", which was held at University of Portsmouth. Dr. Marwan Omar chaired the workshop "Quantum-enhanced Machine Learning: Bridging Classical Data Science with Quantum Computing", which was held at Illinois Institute of Technology. Dr. Ammar Alazab chaired the workshop "Privacy-Preserving Intrusion Detection: Empowering Security with Federated Learning", which was held at Torrens University Australia. Dr. Xinqing Xiao chaired the workshop "Edge Computing and AI based Intelligent Sensing Data Management", which was held at China Agricultural University. Prof. Wang Juan chaired the workshop "Blockchain and Fintech", which was held at Beijing Computer Federation.

We would like to give sincere gratitude to all authors and speakers who have made their contributions to CONF-CDS 2024, editors and reviewers who have guaranteed the quality of papers with their expertise, and the committee members who have devoted themselves to the success of CONF-CDS 2024.

Dr. Elisavet Andrikopoulou

General Chair of Conference Committee

Workshops

Workshop – Portsmouth: Data Visualization Methods in Healthcare

September 12th, 2024 (BST)

School of Computing, University of Portsmouth

Workshop Chair: Dr. Elisavet Andrikopoulou, Senior Lecturer in University of Portsmouth

Workshop – Chicago: Quantum-enhanced Machine Learning: Bridging Classical Data Science with Quantum Computing

September 11th, 2024 (CDT)

ITM Department, Illinois Institute of Technology, USA

Workshop Chair: Dr. Marwan Omar, Associate Professor in Illinois Institute of Technology

Workshop – Melbourne: Privacy-Preserving Intrusion Detection: Empowering Security with Federated Learning

September 23rd, 2024 (UTC +10)

Artificial Intelligence Research Center, Torrens University Australia

Workshop Chair: Dr. Ammar Alazab, Senior Lecturer in Torrens University Australia

Workshop – Beijing: Edge Computing and AI based Intelligent Sensing Data Management

August 12th, 2024 (GMT+8)

College of Engineering, China Agricultural University

Workshop Chair: Dr. Xinqing Xiao, Associate Professor in China Agricultural University

Workshop – Beijing: Blockchain and Fintech

June 2nd, 2024 (GMT+8)

Beijing Computer Federation Blockchain and Digital Finance Specialized Committee

Workshop Chair: Prof. Wang Juan, Director of the Beijing Computer Federation

The 6th International Conference on Computing and Data Science

CONF-CDS 2024

Table of Contents

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241508

Exploring the role of blockchain technology in enhancing data integrity and privacy protection

Xianghui Meng

University of Illinois, Urbana-champaign

xmeng19@illinois.edu

Abstract. This paper systematically builds a theoretical framework for enhancing data integrity and privacy protection by analyzing the fundamentals of blockchain technology and its inherent characteristics, such as decentralization, tamperability, and cryptographic algorithms. The study empirically examines the transparency and anonymity balance mechanism of blockchain in handling sensitive data using case studies and simulation experiments, and at the same time, designs smart contracts to automatically implement data protection strategies for different types of data security threats.

Keywords: blockchain, network security, privacy technology, network security management.

1. Introduction

With the rapid pace of digitization, maintaining data integrity and protecting privacy have emerged as significant challenges within the information technology landscape. The frequent incidents of data breaches and privacy violations pose severe threats to individuals, businesses, and even national security [1]. Blockchain technology, characterized by its inherent resistance to modification, decentralization, and cryptographic security, offers innovative solutions to these challenges [2].

Blockchain technology assures that once data is recorded on its ledger, it cannot be altered covertly, thus safeguarding the data's authenticity and enabling traceability [3]. This immutable nature of blockchain is crucial for the verification and integrity of data. Moreover, blockchain's decentralized architecture reduces the risk of centralized control, thereby enhancing the system's resilience against attacks [4]. In terms of privacy, the application of advanced encryption techniques within blockchain ensures the secure transmission and storage of data, allowing only authorized users access to sensitive information, which is vital for protecting user privacy [5][6].

However, despite the theoretical benefits, blockchain faces practical challenges such as scalability, performance, and integration with existing systems. These issues must be addressed to fully leverage blockchain in safeguarding data integrity and privacy [7][8]. This study aims to explore how blockchain can overcome these limitations to enhance secure data storage, processing, and transmission. It also seeks to propose practical solutions to facilitate the broader application of blockchain in data integrity and privacy protection sectors.

This paper will provide a detailed analysis of blockchain's mechanisms and principles related to data integrity and privacy protection, evaluate its effectiveness in real-world scenarios, and explore potential

DOI: 10.54254/2755-2721/87/20241508

integrations with other cutting-edge technologies like artificial intelligence and the Internet of Things (IoT) to further enhance data security [9]-[14].

2. The effect of blockchain technology application in real scenarios

2.1. Application in Financial Transaction Scene

In financial transaction scenarios, the application of blockchain technology demonstrates its excellent potential in data integrity and privacy protection. The traditional financial transaction system relies on centralized intermediaries, which not only increases transaction costs, but also is vulnerable to single point of failure and fraud. In contrast, the distributed nature of the blockchain allows each transaction record to be replicated and stored on multiple nodes of the network, creating a tamper-proof public ledger (L) that ensures data integrity and transparency.

The formula:

$$\mathrm{L} = \sum_{i=1}^{n} Ti, \tag{1}$$

In this context, T_i represents the *i-th* transaction record, and n is the total number of transaction records. This decentralized architecture eliminates the need for trust in third parties, reducing the time and cost of transaction verification. In terms of privacy protection, blockchain technologies like Zero-Knowledge Proofs (ZKP) allow one party

(the prover P) to demonstrate to another party (the verifier V) that they possess certain information (e.g., ownership of specific funds) without revealing the information itself:

$$\mathrm{P} \rightarrow \mathrm{V} : (\mathrm{proof}|\mathrm{ZKP}), \tag{2}$$

Such interactions ensure the anonymity and privacy of transactions while preventing double-spending attacks and other fraudulent activities.

Furthermore, the implementation of Smart Contracts (SC) on the blockchain further enhances the security and automation of financial transactions:

$$\mathrm{SC}(c, t_1, t_2, \ldots, t_n) \rightarrow \mathrm{Execution}, \tag{3}$$

In this context,c represents the contract conditions,t_i is the events that trigger the execution of the contract. Smart contracts automate predefined rules, reducing the need for human intervention and lowering the risk of breach.

In summary, blockchain technology enhances the integrity and privacy of data in financial transactions through its distributed ledger, zero-knowledge proofs, and smart contracts, bringing revolutionary changes to the financial industry. However, despite these advancements, attention must still be paid to the scalability, energy consumption, and regulatory adaptability of blockchain technology to promote its application in broader fields.

2.2. Application in Medical Record Scenarios

In the context of medical records, the application of blockchain technology demonstrates its exceptional potential for data integrity and privacy protection. Traditional medical information systems often face issues like data silos, patient privacy breaches, and tampering risks. The distributed nature of blockchain, its immutable ledger, and smart contracts provide innovative solutions to these problems.

Firstly, through a decentralized network structure, blockchain technology enables medical data to be stored not in a single institution but distributed across various network nodes. This method of distributed storage (Table 1) reduces the risk of single points of failure, enhancing data availability and reliability. Each block contains the hash value of the previous block, forming a continuous chain-like structure. Any modifications to existing data will result in changes in the hash values of subsequent blocks, thereby facilitating the detection of tampering.

DOI: 10.54254/2755-2721/87/20241508

Table 1. Blockchain Distributed Storage Schematic

Block	Data	Previous Block Hash
1	Patient A's Record	-
2	Update to Patient A's Record	Hash of Block 1
3	Additional Data	Hash of Block 2
...	...	...
n	Latest Update	Hash of Block (n-1)

2.3. Application in Supply Chain Management

In the complex and dynamic field of supply chain management, the application of blockchain technology demonstrates its unique advantages. Traditional supply chain management systems are often limited by information silos, lack of transparency, and trust issues, leading to inefficiency and increased potential risks. Blockchain technology, with its distributed ledger and smart contracts, provides innovative solutions to these problems (Table 1).

In practical applications, blockchain technology significantly enhances the transparency of the supply chain. For example, by recording every transaction on the blockchain, all participants can view the status of goods in real-time, from production to delivery, making the entire process transparent and reducing the possibility of fraud (1-4).

Table 2. Core Advantages of Blockchain Technology in Supply Chain Management

Advantage	Description
Decentralization	Eliminates single points of failure, enhancing system robustness
Transparency	All transaction records are publicly accessible, enhancing trust
Immutability	Once data is recorded, it cannot be altered, ensuring data integrity
Smart Contract	Automatically executes contractual terms, reducing operational errors and disputes
Traceability	Real-time tracking of goods, ensuring traceability from source, improving efficiency

$$T = \sum_{i=1}^{n} T_i, \tag{4}$$

In this context, T represents the overall transparency of the supply chain, while *Ti* indicates the transparency of the *i -th link in the chain.*

Additionally, smart contracts automatically execute contract terms, reducing manual intervention and minimizing operational errors and disputes. When preset conditions are met, the contract automatically executes actions such as payment and delivery, enhancing transaction efficiency (1-5):

$$E = f(C, V), \tag{5}$$

In this context,E represents efficiency improvement,C stands for smart contracts,V denotes the verification and execution process.

In summary, the application of blockchain technology in supply chain management not only enhances data integrity but also effectively protects the privacy of participants. By increasing transparency and automating processes, it significantly improves the operational efficiency and security of the supply chain. However, despite these significant achievements, attention must still be given to the standardization of technology, regulatory adaptability, and the costs of large-scale deployment to promote the widespread application of blockchain in supply chain management.

DOI: 10.54254/2755-2721/87/20241508

3. Comparative Analysis with Traditional Network Security Technologies

In exploring the application of blockchain technology for data integrity and privacy protection, we focuses on the comparative analysis between blockchain technology and traditional network security technologies. Blockchain technology, with its core features of a distributed ledger and cryptographic algorithms, has brought revolutionary changes to network security. Compared to traditional centralized storage and authentication methods, blockchain's decentralized architecture eliminates single points of failure and enhances data immutability, thereby showing significant advantages in protecting data integrity.

Traditional network security technologies rely on firewalls and access control lists to prevent unauthorized access, but these measures are often inadequate in the face of internal threats or advanced persistent threats. In contrast, blockchain's smart contracts provide automatically executed rules, ensuring that only transactions that meet preset conditions are executed, greatly enhancing the security management efficiency of network resources. Additionally, blockchain's cryptographic algorithms, such as Elliptic Curve Cryptography (ECC) and hash functions (H(x)), ensure the privacy of data transmission and storage, making it difficult for intercepted data to be easily decrypted or tampered with.

However, blockchain is not a panacea. It faces challenges such as scalability issues and energy consumption. For example, the transaction processing speed of the Bitcoin network is much lower than that of credit card networks, and it consumes a significant amount of energy. Moreover, although anonymity is a notable feature of blockchain, the maturity of technologies like zero-knowledge proofs still needs to be improved to fully balance privacy protection with anti-money laundering and anti-fraud requirements.

In conclusion, blockchain technology demonstrates strong potential in data integrity and privacy protection, but it also needs to address the limitations of existing technologies for broader and deeper application. Through continuous technological innovation and optimization, blockchain is expected to become an indispensable pillar in the field of network security, complementing traditional technologies to build a more secure and trustworthy network environment.

4. Conclusion

In this research, we have explored in-depth the core role of blockchain technology in data integrity and privacy protection, revealing its potential as the future infrastructure for information security. Our findings are summarized as follows:

Firstly, we have shown how blockchain's distributed nature ensures data immutability. Through the hash-linked block structure (2-1), each block contains the hash value of the previous block, forming a chain. Any modification to historical data will cause changes in subsequent block hash values, which are quickly discovered by nodes in the network.

$$H_i = H\left(data_i \| H_{i-1}\right) \tag{6}$$

Secondly, the introduction of smart contracts has enhanced the transparency and automation of data processing. These self-executing contracts (2-2) define rules and conditions that automatically execute when preset conditions are met, reducing human intervention and enhancing the security of data exchange.

$$f\left(state, input\right) \rightarrow \left(output, new\_state\right) \tag{7}$$

Furthermore, we have explored how Zero-Knowledge Proofs (ZKP) can verify the authenticity of data without disclosing the original information. ZKP allows one party (the prover) to prove to another party (the verifier) that they know certain information without revealing the information itself, thus finding a balance between privacy protection and verification (2-3).

$$P \vdash_{\Sigma} K : \text{Proof} \tag{8}$$

Despite these advancements, blockchain technology still faces challenges related to scalability, energy consumption, and regulatory adaptability. Future research should focus on optimizing consensus

DOI: 10.54254/2755-2721/87/20241508

algorithms, exploring more energy-efficient solutions, and aligning with existing regulatory frameworks to promote the widespread application of blockchain in the fields of data integrity and privacy protection.

In summary, blockchain technology has not only revolutionized the way data is managed and protected but has also laid the foundation for building a more secure and transparent information environment. As the technology matures, we look forward to blockchain demonstrating its unique value in more areas, providing strong support for the global digital transformation.

References

[1] Zhang, X. (2023). Applications of Artificial Intelligence in Cybersecurity. Wireless Internet Technology, 20(06), 29-35.

[2] Lin, H. Y., Wang, J., Niu, D., et al. (2024). Blockchain-driven framework for construction waste recycling and reuse. Journal of Building Engineering, 89109355.

[3] Zhu, L., Jiang, H., Zhu, J., et al. (2024). Hardware method for zero optical path difference position detection of FTIR spectrometer. Measurement: Sensors, 33101153.

[4] Endace. (2020). Endace Wins Big in Cyber Defense Magazine and Info Security Products Guide Awards; EndaceProbe Analytics Platform receives ten awards including Best Security Hardware, Best Packet Capture Product, Most Innovative Security Hardware, and Best Network Security and Management. M2 Presswire.

[5] Dong, Y., & Han, Q.-L. (2019). Guest Editorial Special Issue on New Trends in Energy Internet: Artificial Intelligence-Based Control, Network Security, and Management. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49(8).

[6] IEEE. (2019). Special Issue on New Trends in Energy Internet: Artificial intelligence-Based Control, Network Security and Management. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49(1).

[7] Cai, Z., Hu, C., Zheng, K., Xu, Y., & Fu, Q. (2018). Network Security and Management in SDN. Security and Communication Networks, 2018(2018).

[8] Cyberoam Technologies Pvt. Ltd. (2015). Patent Issued for Identity and Policy-Based Network Security and Management System and Method. Journal of Engineering.

[9] Zubaydi, H. D., Varga, P., & Molnár, S. (2023). Leveraging Blockchain Technology for Ensuring Security and Privacy Aspects in Internet of Things: A Systematic Literature Review. Sensors, 23(2), 788. https://doi.org/10.3390/s23020788

[10] Taherdoost, H. (2023). Privacy and Security of Blockchain in Healthcare: Applications, Challenges, and Future Perspectives. Sci, 5(4), 41. https://doi.org/10.3390/sci5040041

[11] Zhang, X. (2023). Applications of Artificial Intelligence in Cybersecurity. Wireless Internet Technology, 20(06), 29-35.

[12] Lin, H. Y., Wang, J., Niu, D., et al. (2024). Blockchain-driven framework for construction waste recycling and reuse. Journal of Building Engineering, 89109355.

[13] Zhu, L., Jiang, H., Zhu, J., et al. (2024). Hardware method for zero optical path difference position detection of FTIR spectrometer. Measurement: Sensors, 33101153.

[14] Endace. (2020). Endace Wins Big in Cyber Defense Magazineand Info Security Products Guide Awards; EndaceProbe Analytics Platform receives ten awards including Best Security Hardware, Best Packet Capture Product, Most Innovative Security Hardware, and Best Network Security and Management. M2 Presswire.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241560

Facial recognition - A literature review

Shengdi Wang

University of Sheffield, Sheffield S10 2TN, United Kingdom

shengdi777@gmail.com

Abstract. This paper analyses the main technologies for face recognition, a critical biometric tool for identity verification and security across various sectors. A comprehensive overview of traditional and modern facial recognition technologies will be provided, examining their key features such as age, pose, and illumination. The study discusses the evolution and current state of facial recognition, highlighting significant advancements and applications in recent years. The objective is to offer a detailed understanding of how these technologies function and their implications for security and identity verification.

Keywords: face recognition, biometrics, neural networks, applications.

1. Introduction

In recent years, facial recognition technology has significantly advanced[1], becoming an integral part of numerous security and identification systems around the world. Utilised for a variety of applications ranging from law enforcement to personal device security, this technology leverages unique facial features to identify or verify individuals. As a non-intrusive and user-friendly biometric solution, it is favoured in many public and private sectors. The reason why facial recognition technology can develop so rapidly is that it combines many factors: advanced active development of algorithms[5], the availability of a large database of facial images and plenty of methods for evaluating the performance of face recognition algorithms. However, facial recognition technology is facing significant challenges, including variable environmental conditions, ethical concerns, and the need for improved accuracy and privacy safeguards.

Biometric identification consists of determining the identity of a person. Biometrics can be divided into two types – behavioural and physiological.[3]The swift advancement of mobile technology has enabled the incorporation of biometric sensors into smart devices.[4]Physiological biometrics are based on unique physical characteristics which vary from individual to individual, such as fingerprints, iris patterns, facial features and hand geometry. Behavioural biometrics focus on unique patterns in personal activities and behaviours including voice, signature, odour and keystroke dynamics. Although the study of some physiological and behavioural biometrics like ear and nose structure or keystroke dynamics is still in early development stages, each biometric method offers distinct advantages and drawbacks. For example, although iris recognition has high accuracy, it is not cost-effective. Also fingerprints are easily collected but might not work well with uncooperative subjects.

In the realm of security and personal identification, biometrics is used to identify individuals based on their physiological or behavioural characteristics. These methods have been foundational in various applications ranging from secure access control systems to personal device security and law enforcement.

DOI: 10.54254/2755-2721/87/20241560

As technology is advancing, biometric systems have increasingly integrated into daily life, enhancing security protocols but also raising important privacy and ethical considerations. The choice of biometric technique which ranges from facial recognition to fingerprint scanning depends on the specific needs of the application, the required level of security, cost considerations and the acceptability of the technology to users. There are several aspects of the development of facial recognition.

Face recognition algorithms fall into two categories: fully automatic and partially automatic. Fully automatic algorithms handle the entire process independently, from detecting and normalising the face to identifying the individual by comparing features against a database. Partially automatic algorithms need additional data, such as the coordinates of key facial landmarks, to aid in normalisation and identification. The choice between these algorithms depends on automation needs, accuracy, and available computational resources.

Face recognition technology is also classified based on image orientation into frontal, profile, and view-tolerant recognition. Frontal recognition requires direct camera facing and is used in controlled environments like access control systems. Profile recognition handles side views and is suitable for surveillance where direct facing is not possible. View-tolerant recognition can identify faces from various angles, making it ideal for dynamic environments like crowded public spaces. These types enhance biometric systems' versatility, catering to applications from secure access to public safety monitoring.

- Pose

The variability in a person's pose—whether they are facing forward, looking to the side, or tilting their head—poses a substantial challenge for facial recognition systems[2]. A change in head position can alter the appearance of facial features in ways that are not always predictable, making consistent recognition difficult. This issue has been a focal point in facial recognition research for decades. Advances in 3D modelling and multi-angle recognition have been developed to address these challenges. Techniques like view-tolerant recognition algorithms are designed to handle images captured from various angles, but there is still considerable work to be done to perfect these methods across all potential use cases.

- Illumination

Lighting conditions play a critical role in the performance of facial recognition systems[3]. Variations in lighting can create shadows, highlight certain facial features, or obscure others, leading to inconsistent inputs for recognition algorithms. This variability can significantly degrade the accuracy of facial recognition. Researchers are actively exploring solutions to make facial recognition systems more robust against changes in lighting. Techniques such as using infrared illumination or developing algorithms that can normalise lighting conditions in images are among the approaches being considered to mitigate the effects of variable lighting.

- Age

As people age, facial features change significantly, posing challenges for facial recognition systems. Research is ongoing to develop dynamic algorithms and machine learning models that adapt to these changes over time.

Despite advances, two major limitations remain. Firstly, no system can fully handle all facial variations like pose, illumination, expression, and age. Secondly, these systems perform better with more training images, but obtaining extensive datasets is often limited by privacy, logistics, or rare conditions, impacting their effectiveness in diverse environments.

DOI: 10.54254/2755-2721/87/20241560

2. Traditional facial recognition technologies

2.1. Eigenface

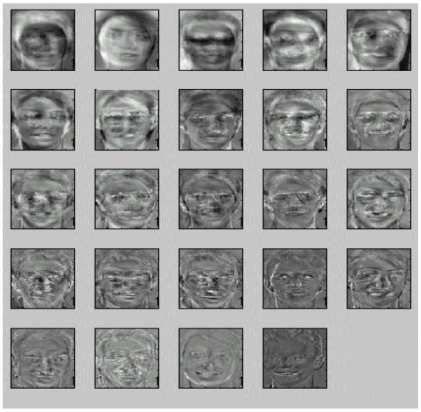

Figure 1. Eigenfaces for sample faces[8]

Eigenface, a key method in facial recognition, relies on Principal Component Analysis (PCA) to reduce image complexity and transform faces into eigenfaces, capturing main variations in features. Each face is represented as a weighted combination of these eigenfaces, with recognition achieved by comparing feature vectors derived from them.

Eigenface is efficient, compressing data significantly and allowing rapid processing during training. It has achieved varied accuracy, with a database of 2,500 images showing correct classifications at rates of 96%, 85%, and 64% for lighting, orientation, and size variations, respectively. However, it requires consistent image quality and conditions, performing poorly with significant variations in lighting, pose, and scale. Faces must be aligned similarly for effective recognition.

The method also struggles with ageing and facial expressions. Adaptations like eigenfeatures target specific components (e.g., eyes, nose) to improve sensitivity to appearance changes. Combining face recognition with other biometrics, such as ear measurements, has significantly improved recognition rates; for example, combining ear and face data increased the rate from 70.5% to 90.9%.

In summary, eigenface is suitable for controlled environments where conditions are standardised, offering a fast and straightforward technique, ideal when speed and simplicity are prioritised over high precision.

2.2. Neural networks

Neural networks have significantly advanced the field of facial recognition by leveraging their inherent non-linearity[9], which enhances the efficiency of the feature extraction process beyond what linear methods like the Karhunen-Loève can achieve. One of the earliest applications of artificial neural networks (ANNs) in facial recognition involved the use of a single-layer adaptive network called WIZARD, which set the foundation for more complex systems like multilayer perceptrons and convolutional neural networks (CNNs). These networks handle tasks from basic face detection to more complex face verification with high accuracy, often incorporating innovative structures such as multi-resolution pyramids and hybrid systems that combine local image sampling with self-organising maps.

A critical advantage of neural networks is their ability to adapt to the variability in facial images through structures that allow for the dimensional reduction of data and partial invariance to changes in translation, rotation, scale, and deformation. For example, hybrid networks reported recognition accuracies as high as 96.2% on databases like the ORL, with 400 images of 40 individuals. However, neural networks are not without challenges; they require extensive training times, with some models taking up to four hours to train, though classification can be completed in under half a second.

DOI: 10.54254/2755-2721/87/20241560

2.3. Graph matching

Graph matching is a sophisticated approach in face recognition, using dynamic link structures and elastic graph matching to handle variations in orientation and expression. Each face is represented by a graph with nodes at specific facial landmarks and edges capturing geometric distances. Features like Gabor filter responses characterise each node.

Elastic Bunch Graph Matching (EBGM) is a notable implementation, organising graphs into a face bunch graph for efficient comparisons. This method handles nonlinear variations such as illumination, pose, and expression, achieving recognition rates up to 98%. It demonstrated high rotation invariance, with success rates of 86.5% and 66.4% for 15 and 30-degree rotations, respectively.

However, graph matching is computationally intensive; comparing 87 objects took about 25 seconds on a system with 23 transputers. It also requires high-resolution images to accurately localise landmarks, posing challenges in scenarios like surveillance with distant or lower-resolution captures. Despite these challenges, graph matching is powerful for managing precise rotational and expressive variability in face recognition.

2.4. 3D morphable model

Figure 2. Different light directions have influence on 3D model fitting[4]

The 3D morphable model in facial recognition uses a vector space representation of faces, combining shape and texture vectors to depict human faces realistically. By fitting these 3D models to images, the system operates under two paradigms: using model coefficients to capture intrinsic face features and creating synthetic views for viewpoint-dependent recognition.

This method incorporates deformable 3D models and computer graphics to estimate 3D shape, texture, and scene parameters like illumination and projection from a single image. It handles non-Lambertian reflections, specular reflections, and cast shadows, automatically adjusting for head position, camera focal length, and illumination. An initialization procedure using six to eight facial points enhances model setup.

Empirical results show the model's efficacy, with 95% accuracy on the CMU-PIE database using side-view galleries and 95.9% on the FERET set with frontal views. Despite requiring high-quality 3D scans and being computationally intensive, the 3D morphable model offers high-precision recognition across diverse conditions, marking a significant advancement in facial recognition.

DOI: 10.54254/2755-2721/87/20241560

3. Modern facial recognition technologies

3.1. Line Edge Map(LEM)

Figure 3. An illustration of a face LEM[12]

The Line Edge Map (LEM) approach in facial recognition leverages edge information, which is less sensitive to illumination variations. LEM uses edge maps to capture facial feature boundaries, maintaining consistent visibility despite lighting changes. By converting edge maps into line segments through polygonal line fitting, LEM reduces model complexity, storing only segment endpoints for a simplified face representation.

The process involves thinning the edge map for precise line fitting, creating an efficient data structure. LEM has shown superior performance, achieving perfect or near-perfect identification rates of 100% and 96.43% on specific databases. It matches the eigenface method's performance under ideal conditions and excels in handling slight appearance and size variations.

However, LEM is sensitive to pose and significant facial expression changes, making it less effective in dynamic environments. Despite this, LEM is a robust, storage-efficient solution for facial recognition, particularly suitable for applications with variable lighting where edge detail is crucial for identity verification.

Table 1. Performance Comparison on the AR Database [11]

Method	Recognition rate
LEM	96.43%
Eigenface (20-eigenvectors)	55.36%
Eigenface (60-eigenvectors)	71.43%
Eigenface (112-eigenvectors)	78.57%

According to the table, it can be found that the recognition rate of LEM is the best, compared to other methods.

3.2. Support Vector Machine(SVM)

Support Vector Machines (SVM) are highly effective for face recognition, leveraging an Optimal Separating Hyperplane (OSH) to maximise the margin between classes, minimising misclassification risk. This is achieved through Structural Risk Minimization (SRM), optimising both training and generalisation errors, making SVM suitable for limited training samples.

SVMs handle high-dimensional data well, ideal for face recognition with numerous input features. Kernels allow SVMs to operate in transformed feature spaces, managing complex, nonlinear relationships. This enables effective modelling of facial image similarities and dissimilarities for verification and identification tasks.

DOI: 10.54254/2755-2721/87/20241560

Challenges include computational intensity, particularly with large datasets and complex Quadratic Programming (QP) during training. Decomposition algorithms help mitigate this, but performance can degrade with many classes (individual faces).

Empirical results highlight SVM robustness in face recognition, achieving 92% accuracy with edge maps and outperforming traditional methods like eigenfaces under varying conditions. For example, SVMs achieved an 8.79% error rate compared to 15.14% with Nearest Center Classification (NCC) using eigenfaces. SVMs also excel in multi-view face detection, with over 90% recognition accuracy and a 95% detection rate in video sequences.

Overall, SVMs offer high accuracy and robustness in face recognition, despite some computational and scalability challenges.

3.3. Multiple Classifier Systems(MCSs)

Multiple Classifier Systems (MCSs) enhance face recognition by integrating outputs from various classifiers, improving accuracy and robustness. This approach leverages the diverse strengths of classifiers like Learning Vector Quantization (LVQ), Radial Basis Function (RBF) neural networks, Eigenfaces, Fisherfaces, Support Vector Machines (SVM), and Elastic Graph Matching (EGM). By combining different classifiers, MCSs reduce uniform errors and handle variations in pose, expression, and illumination better than single classifiers.

Hybrid methods using both holistic and feature-based analyses, like the Markov Random Field (MRF) model, also performed well, with recognition rates of 96.11% on Yale and 86.95% onORL.

Designing and training MCSs can be complex and computationally expensive due to the need to integrate multiple classifiers effectively. However, their superior accuracy and adaptability to varied data conditions make MCSs a powerful tool in advanced face recognition systems.

4. Application

- Security access control

Facial recognition technology is increasingly used in high-security access control systems, such as the Chui doorbell by Trueface.ai, which employs deep learning to distinguish real faces from photos, preventing fraud. This system scans a face and compares it to a database of authorised individuals, granting access if a match is found. While enhancing security and convenience by eliminating the need for physical keys, it raises privacy concerns over sensitive biometric data and can be affected by poor lighting or changes in appearance. Despite these issues, facial recognition remains a valuable tool for secure access control.

- Surveillance systems

Surveillance systems, especially those using facial recognition technology, are increasingly deployed across various sectors for enhanced security and monitoring. These systems utilise CCTV cameras installed at strategic locations to capture video footage, which is then processed to identify individuals based on facial recognition.

The principle behind these systems involves capturing live video feeds, extracting faces, and matching them against a database of known individuals to identify potential offenders or track customer behaviour.

However, while facial recognition in surveillance offers considerable benefits such as enhanced security and loss prevention, it also comes with drawbacks. Privacy concerns are significant, as there is the potential for misuse of personal biometric data. Furthermore, the accuracy of these systems can be affected by various factors including poor lighting, obstructions, or changes in appearance. Despite these challenges, facial recognition remains a powerful tool in modern surveillance systems, providing a mix of proactive security and substantial utility in public safety operations.

- General identity verification

DOI: 10.54254/2755-2721/87/20241560

General identity verification systems increasingly incorporate facial recognition technology to validate personal identities using important documents such as national identification cards and passports. The core principle involves capturing a facial image of the individual, which is then digitised and stored as part of their official identity record. During verification, the stored image is compared with a live capture or another submitted image to confirm the person's identity.

This method provides a quick and efficient way of confirming identities, enhancing security for various administrative and legal processes. However, it also raises privacy concerns due to the storage and handling of personal biometric data. Additionally, the accuracy of facial recognition can be compromised by poor image quality or changes in a person's appearance over time. Despite these potential drawbacks, facial recognition in identity verification remains a valuable tool for enhancing security and streamlining identification processes.

5. Conclusion

Facial recognition technology is widely used in security, surveillance, and personal device access. It strengthens access control by verifying identities through deep learning algorithms, and in surveillance, it helps monitor and identify individuals in places like banks and malls, reducing criminal activities. For identity verification, it's used in passports and national IDs, matching individuals with biometric data in government databases. On personal devices, facial recognition offers a quick alternative to PINs and passwords, despite challenges with environmental variations and security vulnerabilities.

However, this technology raises privacy concerns due to the risks of data breaches and unauthorised surveillance. Its effectiveness can also be affected by poor lighting, changes in appearance, and capture angles. Despite these issues, facial recognition continues to evolve, enhancing security and user experience in various applications.

References

[1] Zhao, W., Chellappa, R., Phillips, P. J., & Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM computing surveys (CSUR)*, *35*(4), 399-458.

[2] Adini, Y., Moses, Y., & Ullman, S. (1997). Face recognition: The problem of compensating for changes in illumination direction. *IEEE Transactions on pattern analysis and machine intelligence*, *19*(7), 721-732.

[3] Jain, A. K., & Li, S. Z. (2011). *Handbook of face recognition* (Vol. 1, p. 699). New York: springer.

[4] Kaur, P., Krishan, K., Sharma, S. K., & Kanchan, T. (2020). Facial-recognition algorithms: A literature review. *Medicine, Science and the Law*, *60*(2), 131-139.

[5] Zhou, S. K., & Chellappa, R. (2005). Image-based face recognition under illumination and pose variations. *JOSA A*, *22*(2), 217-229.

[6] Partridge, D., & Griffith, N. (2002). Multiple classifier systems: Software engineered, automatically modular leading to a taxonomic overview. *Pattern Analysis & Applications*, *5*, 180-188.

[7] Lanitis, A., Taylor, C. J., & Cootes, T. F. (2002). Toward automatic simulation of ageing effects on face images. *IEEE Transactions on pattern Analysis and machine Intelligence*, *24*(4), 442-455.

[8] Ellavarason, E., Guest, R., Deravi, F., Sanchez-Riello, R., & Corsetti, B. (2020). Touch-dynamics based behavioural biometrics on mobile devices–a review from a usability and performance perspective. *ACM Computing Surveys (CSUR)*, *53*(6), 1-36.

[9] Kshirsagar, V. P., Baviskar, M. R., & Gaikwad, M. E. (2011, March). Face recognition using Eigenfaces. In *2011 3rd International Conference on Computer Research and Development* (Vol. 2, pp. 302-306). IEEE.

[10] Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience*, *3*(1), 71-86.

[11] Gao, Y., & Leung, M.K. (2002). Face Recognition Using Line Edge Map. *IEEE Trans. Pattern Anal. Mach. Intell., 24*, 764-779.

DOI: 10.54254/2755-2721/87/20241560

[12] Yongsheng Gao and M. K. H. Leung, "Face recognition using line edge map," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 6, pp. 764-779, June 2002.

[13] Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1701-1708).

[14] Ratcliffe, J. (2006). *Video surveillance of public places*. Washington, DC: US Department of Justice, Office of Community Oriented Policing Services.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241569

The survey and discussion of research on heart disease prediction based on Apache Spark

Junyu Hu

The Department of Artificial Intelligence, South China Normal University, Foshan, 528225, China

20224001049@m.scnu.edu.cn

Abstract. Heart is the most important part of the human body, diseases that are related to heart cause a huge thread to human health. In this paper, methods that applied Apache Spark to heart disease related works would be shown and discussed in order to classify these methods and make a conclusion about the innovations and shortcomings of these works. These works are defined into two categories: the ones that adopted traditional machine learning method and the ones that used deep learning methods. By classifying these works into two types, commonalities and similar innovative approaches in the same category of methods can be better observed and summarized, facilitating a clearer comparison of the similarities and differences in the innovative focuses among similar yet distinct methods. By doing so, conclusions were made to show that apart from enhancing operational efficiency and reliability of models for diagnosing and treating heart diseases, current research utilizing Apache Spark in this field also identifies areas for improvement such as expanding sample data representation, speeding up data processing, and addressing concept drift issues with proposed solutions. By addressing these challenges, researchers aim to optimize existing methods using Apache Spark and advanced data analytics techniques to combat heart diseases.

Keywords: Apache Spark, machine learning, deep learning, monitoring and prediction of heart disease, electrocardiogram analysis.

1. Introduction

As one of the most significant organs of the human body, the heart is like an engine of a car, it delivers oxygen and other nutrition that other organs need to keep functioning to every corner of the body through blood. Heart disease is hard to be diagnosed when it doesn't seizure while its seizures suddenly without an obvious omen, making it hard to prevent. Therefore, diseases that are related to hearts have now become the number one killer of human health: About 695,000 people died from heart disease in 2021—that's 1 in every 5 deaths. Heart disease costs about $239.9 billion each year from 2018 to 2019 [1].

Luckily, there have been new ways to prevent heart disease. By applying wearable medical devices and other devices, medical workers are now able to monitor patients' real-time situation of their hearts, which allows doctors to react timely once heart disease occurs. This method will surely help in the problems that the suddenness of heart disease brings, but it costs too much labor costs for a medical worker to keep monitoring the patient's heart. Which is why applying artificial intelligence into

DOI: 10.54254/2755-2721/87/20241569

monitoring and predicting heart disease is brought up. To achieve this goal, the program should be able to measure real-time heart rate with low latency, and it should have the ability to resist interference since there will be inevitable interference in real life usage. Moreover, the huge amount of data for real-time monitoring demands a program that has a strong ability to deal with big data.

Apache Spark, as a fast and universal computing engine designed specifically for large-scale data processing, meets these requests. Apart from that, it has been used in many other practical fields such as text mining、human face recognition and so on, thus proving its practicality. Therefore, many researchers have applied Apache Spark on heart disease area and have focused on different aspects.

In this article, the following works would be mentioned and analyzed. Ilbeigipour et al. [2] presents the implementation of a machine learning pipeline for real-time heart arrhythmia detection using a structured streaming module built on the open-source Apache Spark platform. The study evaluates the impact of employing this new module on classification performance metrics and the latency rate of heart arrhythmia detection. Alarsan et al. [3] proposes an Electrocardiogram (ECG) classification method using machine learning on Apache Spark, addressing irregularities in ECG signals. Implemented in Scala with MLlib, it achieved high accuracy using algorithms like Gradient-Boosted Trees and Random Forests, leveraging Spark's scalability for processing large datasets. The approach was validated on MIT-BIH Arrhythmia and Supraventricular Arrhythmia databases, demonstrating efficient ECG signal classification. Carnevale et al. [4] proposed a tool based on Apache Spark and the Menard algorithm for handling this electrocardiogram data. To validate their solution, they conducted a series of experiments, implementing the algorithm to detect heart diseases. The experimental results demonstrated the superiority of their approach in terms of performance. Abdullah, et al. [5] proposes a real-time heart rate prediction system based on Apache Spark. By integrating Apache Kafka and Apache Spark, the online phase predicts heart rate in advance using the best model, aiding healthcare providers and patients in real-time avoidance of heart rate risks. With the hope of providing categorized references for relevant researchers, this article will provide a review of the relevant studies on Spark in the field of cardiovascular diseases, aiming to offer insights for fellow professionals.

This paper will investigate the following aspects: firstly, it will present research related to heart disease using Spark with machine learning and deep learning. Subsequently, the paper will analyze and discuss the mentioned research efforts to identify their innovations and strengths. Following this, a summary of these works will be provided. Finally, the paper will conclude with a review of the application of Apache Spark in the field of heart disease.

2. Methods

2.1. Introduction of Spark

Spark is a general-purpose big data processing framework. Similar to traditional big data technologies like Hadoop's MapReduce, Hive engine, and the Storm real-time streaming engine, Spark encompasses various common computing frameworks in the big data field. The working mechanism involves a master-slave architecture where the master, called the driver, coordinates the execution of tasks on worker nodes. When a user submits an application, the driver obtains resources from the cluster manager and then divides the tasks into smaller units of work called tasks. These tasks are then dispatched to the worker nodes for parallel execution. Spark employs in-memory computation, which enhances processing speed by caching intermediate results. It operates on Resilient Distributed Datasets (RDDs), in-memory data structures allowing fault-tolerant distributed processing of large datasets. The Directed Acyclic Graph (DAG) scheduler optimizes task execution, while the use of lazy evaluation minimizes unnecessary computations. These principles—distributed task execution, in-memory computation, RDDs, and optimized task scheduling—form the basis of Spark's operational framework and underpin its efficiency in processing large-scale data.

DOI: 10.54254/2755-2721/87/20241569

2.2. Traditional machine learning methods

2.2.1. Heart arrhythmia detection

The study developed a real-time pipeline for atrial fibrillation and RBBB (Right Bundle Branch Block) arrhythmia detection using ECG (electrocardiogram) signals [2]. It employed online segmentation and feature extraction, with random forest classification shown in Figure 1 [2]. Data from MIT/BIH database were preprocessed for noise removal, R-peak detection, and feature extraction. An Apache Spark pipeline with Pandas-UDF was implemented for data preprocessing. The pipeline used Spark structured streaming for real-time processing.

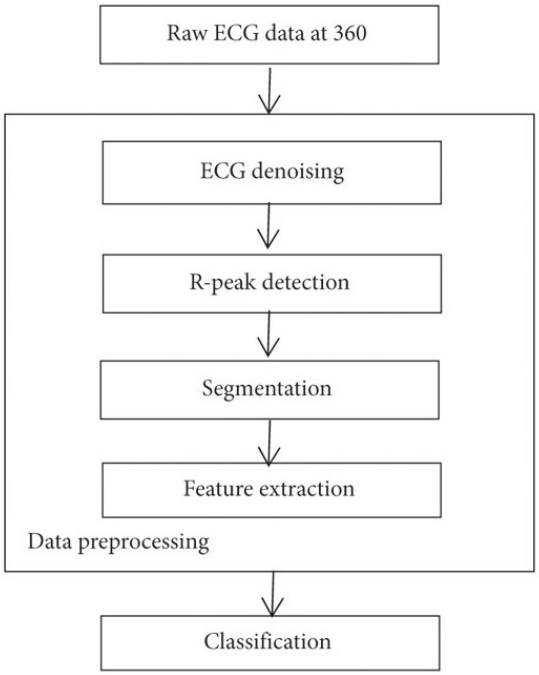

Figure 1. Block diagram of data preprocessing and classification [2].

The classification process involves learning from training tuples to describe class labels and evaluating the model for classification. Various classifiers like decision trees and random forests are trained, leveraging Apache Spark's functionalities for real-time ECG classification. Ilbeigipour et al. innovatively developed an online pipeline integrating preprocessing and classification steps on Spark structured streaming. This approach reduces latency in arrhythmia diagnosis using parallel computing on big data platforms. It also enhances multi-class classification performance and can incorporate static patient data for result reliability.

2.2.2. Heart diseases classification

This study [3] proposes an ECG classification method using Apache Spark, MLlib, and Scala, achieving high-precision signal classification. Evaluated on MIT-BIH Arrhythmia and Supraventricular Arrhythmia databases, it utilizes Discrete Wavelet Transform for feature extraction. Three feature types (Summits, Temporal, Morphological) are categorized. Spark is employed for feature processing.

Due to the large size of the dataset in this case, and the necessity to implement decision trees, random forests, and gradient boosting trees, machine learning algorithms were not easily implementable in Matlab due to performance concerns. Hence, Spark-shell and Scala were utilized on a local host PC. Following the data processing, the authors employed Gradient-Boosted Trees (GBT) and Random Forest (RF) models for machine learning, achieving classification accuracies of 96.75%

DOI: 10.54254/2755-2721/87/20241569

and 97.98%, respectively. The GBT and RF models developed in this study are capable of classifying various types of ECG heartbeats, thus enabling their implementation in CAD ECG systems for fast and reliable diagnosis.

2.2.3. Heart disorder detection

In this paper, Carnevale et al. [4] discuss a tool utilizing the Menard algorithm based on Apache Spark. The core idea of the Menard algorithm is to locate specific peaks (such as QRS peaks) within the ECG signal and determine their positions. Regarding the dataset, the authors utilized the European ST-T database. In the proposed method, the Menard algorithm has been employed for calculating the QRS complex using Apache Spark. During the preprocessing stage, it's necessary to handle multiple samples simultaneously, requiring the organization of a suitable set of samples on a file line because Apache Spark treats each sample as a string RDD. Additionally, Spark distributes the workload across tasks involving the processing of multiple lines of RDDs. The implementation utilizes a set of samples with a duration of 10 seconds, where the only information required for computing the QRS complex is represented by detected peak indices. This scientific work addresses the issue of distributed processing of electrocardiogram (ECG) signals and it resolves the issue of local preprocessing of ECG signals. Apache Spark was utilized as a tool for extensive data preprocessing in this study, enhancing the efficiency of preprocessing.

2.3. Deep learning methods

2.3.1. Prediction for heart rate

In this study, Alharbi et al. [5] proposed a real-time heart rate prediction system for preemptive heart risk avoidance [5] shown in Figure 2. The system comprises two stages: an offline stage and an online stage. The objective of the offline stage is to develop models using various prediction techniques to minimize the root mean square error. In the online stage, Apache Kafka and Apache Spark are employed to predict heart rates in advance based on the best-developed model.

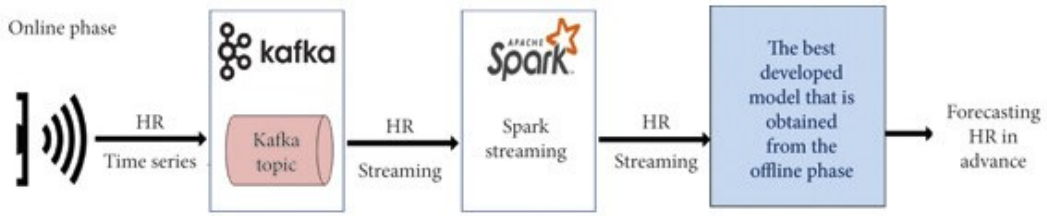

Figure 2. The architecture of HR forecasting system [5].

In terms of dataset, this study utilized the open healthcare dataset called Medical Information Mart for Intensive Care (MIMIC-II). From this dataset, a patient's heart rate time series (univariate dataset) was extracted on a per-minute basis. In the preprocessing stage, the authors first converted the raw data into fixed data, then transformed it into supervised learning format, and finally augmented the data. For model training, four deep learning models were employed for heart rate prediction: Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BI-LSTM), and Gated Recurrent Unit (GRU). These steps belong to the offline stage. In the online stage, Apache Kafka and Apache Spark were utilized as the streaming processing platforms. Apache Spark Streaming API was used to process medical data (i.e., HR streams) from pre-created Kafka topics.

3. Discussion

Although many innovative and effective methods have been proposed shown in the method section, there are still limitations and areas for improvement in the approaches mentioned.

DOI: 10.54254/2755-2721/87/20241569

In the context of Heart Arrhythmia Detection, as noted by the authors, there is room for improvement in enhancing the reliability of patient outcomes by incorporating diverse features into the data. The methods mentioned in the paper also suffer from the high computational complexity associated with newer techniques such as deep neural networks, which have not yet been utilized. Additionally, the issue of concept drift due to variations in data caused by factors like noise remains unaddressed. This phenomenon can impact the effectiveness of new features in online learning and potentially affect the accuracy of results. The methods outlined in the paper still do not adequately handle concept drift in real-world scenarios, highlighting the need for a solution to address this challenge.

In Heart diseases classification, optimizing feature selection and engineering processes can further enhance the performance of classification algorithms. As data volume increases, there is a need for further optimization of algorithms to improve computational efficiency and scalability, particularly for real-time applications. Moreover, since this experiment has not been validated in real-world scenarios, the effectiveness of these models across different datasets or in practical applications remains to be further verified. Lastly, similar to what was mentioned in Heart Arrhythmia Detection, ECG data may experience concept drift due to changes in patient conditions, equipment variations, or other factors, which could impact model performance. Future work could explore better methods to handle concept drift, such as online learning or dynamic model updating.

As for the Heart Disorder Detection, there would be potential challenges in local preprocessing of large files, especially when handling extensive datasets, which could constrain the application's performance. To address this, leveraging a distributed file system for data processing with Spark is recommended to boost efficiency and speed, thereby circumventing single-point bottlenecks. Additionally, while the R-R interval method is commonly used for heart rhythm analysis, it may not comprehensively detect all types of arrhythmias present in complex ECG signals. Therefore, enhancing the analysis by integrating additional signal features and employing deep learning models can effectively identify more intricate arrhythmia patterns. This approach not only improves the accuracy of arrhythmia detection but also expands the scope of analysis to include more nuanced cardiac abnormalities. By combining distributed computing capabilities with advanced analytical techniques, such as deep learning, the article proposes a robust framework for enhancing the precision and scalability of heart disorder detection systems based on ECG data processing.

In terms of the Prediction for Heart Rate section cited from the paper, there are still opportunities for improvement in the stability of the real-time prediction system. During the online phase, simulated sensor-generated heart rate time-series data are sent to a Kafka topic and then processed through Spark streaming before being fed into the best model. However, this architecture may encounter delays or insufficient processing capacity when dealing with large-scale or high-frequency data. Enhancements could involve optimizing the data transmission and processing architecture, such as introducing higher-performance data streaming systems or scaling up server resources. Additionally, the paper does not address the system's capability to handle concept drift, a challenging issue often encountered in real-world scenarios that could affect model performance. Future work could explore better methods to handle concept drift effectively.

In summary, current applications of Apache Spark in conjunction with heart disease detection can benefit from several improvements. Firstly, increasing data diversity by incorporating diverse features can enhance model reliability [6, 7]. Secondly, addressing concept drift challenges more effectively in real-world applications remains a priority for further optimization [8-10]. Lastly, enhancing the performance of classification algorithms can improve computational efficiency and scalability of the algorithms.

4. Conclusion

In this article, the primary focus lies in exploring the application of Apache Spark within the field of cardiology. It delves into various studies that employ both machine learning and deep learning techniques. These research efforts are predominantly geared towards enhancing the operational

DOI: 10.54254/2755-2721/87/20241569

efficiency and reliability of models used in diagnosing and treating heart diseases, each contributing its own unique innovations. Despite the significant strides made in these areas, there remain several avenues for improvement. For example, there is a pressing need to diversify sample data to better represent diverse patient populations. Additionally, improving the speed of data processing could further streamline diagnostic processes and treatment planning. Moreover, researchers have highlighted the concept drift issues that models encounter in real-world applications, underscoring the need for more robust solutions to maintain model accuracy over time. Looking ahead, future research in this domain could benefit from focusing on these areas for enhancement. By addressing these challenges, researchers can potentially refine existing methodologies and develop more effective tools and models for combating heart diseases using Apache Spark and advanced data analytics techniques.

References

[1] CDC 2024 Heart Disease Facts https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html

[2] Ilbeigipour S et al. 2021 Real-Time Heart Arrhythmia Detection Using Apache Spark Structured Streaming Journal of Healthcare Engineering vol 2021 (1) p 6624829

[3] Alarsan F I & Younes M 2019 Analysis and classification of heart diseases using heartbeat features and machine learning algorithms Journal of big data vol 6 (1) pp 1-15

[4] Carnevale L et al. 2017 Heart disorder detection with menard algorithm on apache spark Service-Oriented and Cloud Computing: 6th IFIP WG 2.14 European Conference ESOCC 2017 (Oslo: Springer International Publishing) p 6

[5] Alharbi A et al. 2021 Real-Time System Prediction for Heart Rate Using Deep Learning and Stream Processing Platforms Complexity vol 2021 (1) p 5535734

[6] Nguyen-Tang T & Arora R 2024 On sample-efficient offline reinforcement learning: Data diversity posterior sampling and beyond Advances in neural information processing systems vol 36 Feb 13

[7] Qiu Y Hui Y Zhao P Cai CH Dai B Dou J Bhattacharya S & Yu J 2024 A novel image expression-driven modeling strategy for coke quality prediction in the smart cokemaking process Energy vol 294 May 1 p 130866

[8] Arora S Rani R & Saxena N 2024 A systematic review on detection and adaptation of concept drift in streaming data using machine learning techniques Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery Mar 19 e1536

[9] Fernando DW & Komninos N 2024 FeSAD ransomware detection framework with machine learning using adaption to concept drift Computers & Security Feb 1 vol 137 p 103629

[10] Ali Abdu NA & Basulaim KO 2024 Machine learning in concept drift detection using statistical measures International Journal of Computers and Applications May 3 vol 46 (5) pp 281-291

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241554

Research on the optimize doctor-patient matching in China

Zhihan Luo

University of Southampton, Southampton, SO15 2YU, United Kingdom

zl21u22@soton.ac.uk

Abstract. In China, patients are given the opportunity to know background information about their doctors, and patients choose their doctors to see. Due to the shortage of medical resources and the uncertainty of the quality of medical services, as well as some external factors, patients will prefer to choose well-known doctors in the hope of getting a better medical experience. In order to match the preferred healthcare resources, they may choose to lie to increase the chance of visiting the doctor. This study adopts a combination of theoretical modeling and algorithmic simulation. Through theoretical analysis, a framework model of doctor-patient matching is established. Doctors with different skills and experience, who select patients according to their conditions and their own preferences in the matching process, as well as patients with different conditions and preferences, who want to receive treatment by matching to a suitable doctor, are clarified. The Deferred Acceptance algorithm is written and operated to simulate the matching process where patients apply to doctors and doctors are screened based on their priorities. Analyze and evaluate the performance of Strategy-Proofness and Pareto Efficiency in matching by iterating the algorithm. In this case, the DA algorithm establishes a stable match between the patient and the physician despite the possibility that the patient may deceive his/her preferences. However, patient behavior may affect the efficiency and fairness of the matching process, highlighting the importance of transparency and integrity of the doctor-patient matching system.

Keywords: Deferred Acceptance, Strategy-Proofness, Pareto Efficiency, Doctor-patient relationship.

1. Introduction

China has experienced several major reforms of its health-care system, which have had a profound impact on the doctor-patient relationship in China. Public hospitals are the backbone of China's medical institutions, but are under pressure from insufficient funding and self-reliance requirements. Although the standard of medical care has been improving and the level of medical services has become more sophisticated, the gap between health demand and supply still exists. High drug prices, difficulties in seeking medical advice, and unreasonable use of medication are problems that have plagued the public's access to medical care in recent years. The people's demand for medical services has grown substantially, and problems such as high prices for existing medicines and overmedication have led to an imbalance between supply and demand, putting pressure on the medical system. It wastes limited medical resources, causes tension between doctors and patients, and affects people's trust in the entire medical system [1]. On top of that, China's medical resources are unevenly distributed, with high-quality medical resources mainly concentrated in big cities and well-known hospitals, resulting in patients' demand for famous doctors far exceeding the supply. When studying the doctor-patient problem, patients will be more likely

DOI: 10.54254/2755-2721/87/20241554

to want to be treated by experienced, qualified and veteran doctors for a better medical experience. Patients can find more experienced doctors by knowing the ability, skill, and preference of the general practitioner through hospital posters and the Internet, as well as by comparing other information about the doctor (e.g., feedback from other patients, the doctor's reputation, and success stories). In order to avoid being treated by a younger doctor, patients may choose to exaggerate their condition to increase their chances of being seen by an experienced doctor. The authors' study showed anecdotal evidence in the mass media that patients often intentionally lie, mislead, and deceive the health care professionals who serve them. In addition to this, WebMD, a health and wellness website, found that nearly 45% of respondents admitted to lying to their doctors, more than 30% lied about diet and exercise, and about 40% lied about following their doctor's treatment plan [2]. Jiang and Yuan believe that this phenomenon may lead to some patients who can only be operated by senior doctors facing long waiting times, leading to deterioration of patients' conditions and affecting the quality and effectiveness of healthcare services [3]. Patients lying to obtain better healthcare resources can increase the burden on the healthcare system, exacerbate the shortage of healthcare resources, and affect the stability and efficiency of the entire healthcare delivery system [4]. Chinese patients can indirectly choose their surgeons by selecting their primary care physicians. Different doctors have different proficiencies and preferences for dealing with different conditions, and if a doctor is assigned to an area in which he does not specialize, his or her specialties will not be fully utilized, and work efficiency will be reduced [3]. Therefore, it is not enough to assign a patient to a physician based on the patient's preference alone; it is necessary to consider both the surgeon's and the patient's preference information, to realize two-way matching between surgeons and patients, and to determine a reasonable and effective physician-patient matching. If only one side of the surgeon's or patient's preference is considered and the other side's needs and preferences are ignored, unstable matching may occur. For example, a particular patient is assigned to a doctor whom he does not like and who is also dissatisfied with the match. Jiang and Yuan showed that in this case, it would be better to have another doctor whose preferences are more closely matched to those of the patient, thus affecting the effectiveness of the existing program [3]. To ameliorate this problem, this study aims to optimize the allocation of healthcare resources by using the Delayed Acceptance DA algorithm from the Stable Marriage Theory to create a stable match between patients and doctors. The effectiveness and fairness of this matching mechanism are also demonstrated using Pareto Efficiency theory.

2. Assumptions and definitions framework mode

Based on the current status quo, it is assumed that in China, the application is provided by the patient to the doctor, but the doctor has no power to reject it. Therefore, we try to apply the DA algorithm to reduce the incentives of patients to lie in order to alleviate the status quo in China. Find a stable match between doctors and patients based on their real needs. To improve the stability and efficiency of the medical service system. Assume in this framework mode:

D is the set of doctors

P is the set of patients

Both patients and doctors have their own preference lists that represent their ordering of doctors and patients, respectively.

Each patient $p \in P$ has a preference list

Each doctor $d \in D$ has a preference list

Patients may be matched to better doctors by exaggerating their conditions. Theory Strategy proofness is an important concept in mechanism design and game theory. A mechanism or algorithm is said to be strategy-independent if, for any participant, honestly reporting its true preferences is always the optimal strategy, regardless of the reports of other participants. [5] In other words, participants obtain the best outcome by reporting their true preferences and have no incentive to obtain a better outcome by misreporting or manipulating preferences. Pathak and Sönmez [6] argue that Strategy proofness is one of the attributes of DA and can be considered as an element of fairness. By encouraging participants to report preferences honestly, it reduces the likelihood of system manipulation and improves the fairness

DOI: 10.54254/2755-2721/87/20241554

and efficiency of matching. In the allocation of medical resources, the use of strategy-independent matching algorithms can effectively reduce the phenomenon of patient exaggeration, optimize the use of medical resources, and improve the efficiency and quality of overall medical services. Second, each patient has its own Strict preference ranking. means that each element (e.g., doctor or patient) in the preference list has a unique ranking and no two elements are considered equivalent. In this ranking, each preference is explicitly above or below all other preferences, forming a linear, non-repeating order. And the patient applies to the doctor according to his or her preferences, and the doctor has the power of rejection through iteration. Considering the preferences expressed by the patient, the doctor temporarily accepts the application of the top-ranked patient.

3. Pareto efficiency

The DA algorithm is effective in ensuring Pareto efficiency in the matching process in cases where patients are likely to exaggerate their conditions in order to obtain a more experienced doctor. Patient exaggeration does not change the final matching result, thus reducing the incentive for patients to exaggerate their conditions.

3.1. Definition of Pareto efficiency

Lionel Robbins accepts Vilfredo Pareto's definition of efficiency, which states that a given allocation is efficient when and only when it is impossible to change it without incurring a loss to some people [7]. Thus, when reallocating resources, only those changes that improve the welfare of some without causing losses to others are considered welfare improvements. In other words, Pareto efficiency is a state of resource allocation in which no improvement can benefit at least one person without harming others.

3.2. The role of Pareto efficiency

The role of Pareto efficiency in this problem is to ensure that no possible re-matching can make some patient-doctor matches better without making other patient-doctor matches worse. In the China case, every doctor and patient have been successfully matched. Both doctors and patients have literally gotten their first choice, and no other doctors or patients have been made worse off as a result, which also reflects Pareto efficiency. It also means that resources are allocated most efficiently, and no individual can get more resources without harming others. The concept of Pareto efficiency is particularly important in doctor-patient matching because it ensures fair distribution and optimal utilization of healthcare resources. This efficiency is achieved through DA algorithms that ensure that the final matching result is fair and efficient. By taking into account the real preferences of both doctors and patients, DA algorithms reduce the incentives for patients to exaggerate their conditions and increase the efficiency of medical resources. Ultimately, each matching outcome is Pareto efficient, i.e., there is no possible reallocation that can make some doctors and patients more satisfied without harming others. This approach not only optimizes the allocation of resources, but also enhances the trust between patients and physicians and improves the stability and efficiency of the overall healthcare delivery system.

4. Deferred Acceptance Algorithm (DA)

The DA algorithm is a commonly used matching algorithm that was originally used to solve the stable marriage problem. In this algorithm, both parties involved in a match apply and accept based on their own list of preferences. In the last few years, the algorithm has appeared in an "iterative" (or sequential) mechanism for matching students with schools and universities on a very large scale [8].

4.1. Model Setting

Assume that patient P = {1,2,3} and doctor D = { A,B,C}.

Patient's preference list: the patient has a strict order of preference for the doctor, e.g., patient 1's preference is A≿B≿C.

Doctor's preference list: doctors have a strict order of preference for patients. For example, doctor A's preference is 1≿2≿3.

DOI: 10.54254/2755-2721/87/20241554

and the priority order is in order according to Figure 1 and 2.

$\succsim_1$	$\succsim_2$	$\succsim_3$
A	B	A
B	C	C
C	A	B

Figure 1. Patient 1's preference

$\succsim_A$	$\succsim_B$	$\succsim_C$
1	2	3
2	3	1
3	1	2

Figure 2. Doctor A's preference

4.2. Steps in detail

1) Initial application: Each patient applies to the doctor they most wish to see based on their preference list.

2) Physician Screening: Each physician ranks the applicants in their own order of preference and rejects the lowest ranked patients who exceed their capacity. Patients who are temporarily retained will continue to be retained by the physician.

3) Recursive Application: In a subsequent step, all rejected patients will reapply based on the next physician on their preference list. Each physician combines the new round of applicants with the previously retained applicants, again sorted by priority, and rejects the lowest-ranked patients who exceed their capacity. Patients who are not rejected will continue to be retained by the physician.

4) Termination condition: the process terminates when no new rejections occur. Each physician is matched with the final retained patients and patients not accepted by any physician are not matched.

4.3. Application to the doctor-patient matching problem regarding the explanation of patient's exaggerated condition

In this example, however, we will hypothesize that Patient 3 lies about wanting to be matched to Doctor A. By iterating the DA algorithm

Patient 3 proposes a match to Doctor A, who rejects it. The preferred patient that should be Doctor A's first choice is Patient 1, and Patient 1 has been matched successfully. In addition to this, Patient 3's second choice is Doctor C. Patient 3 proposes a match to Doctor C, who accepts.

Doctor C's preferred patient is Patient 3, so Doctor C and Patient 3 have been matched successfully.

The patient submits a ranked list of preferences to the doctor and iterates until a stable match is reached, which ensures that no patient has an incentive to deviate from their assigned doctor. Balinski and Sönmez show that the DA algorithm is considered to be the "best" fairness mechanism because it is policy-proof and because it has the advantage of being policy-proof This is because it is strategy-proof and Pareto better than any other fairness mechanism (i.e., it is constrained to be efficient) [9].

4.4. Summary

By using the Deferred Acceptance Algorithm, patients apply to physicians based on their preferences, and physicians screen patients based on their priorities. The algorithm ensures that the final match is stable, and because the Deferred Acceptance Algorithm has Strategy-Proofness, the patient has no incentive to manipulate the match by exaggerating the condition. This matching method takes into account both the patient's preferences and the doctor's priorities, thus optimizing the doctor-patient

DOI: 10.54254/2755-2721/87/20241554

matching process, reducing the patient's incentive to exaggerate his or her condition, and improving the overall efficiency and fairness of healthcare services.

5. Conclusion

In China's healthcare system, it is a common phenomenon that patients exaggerate their conditions in order to be matched with more experienced doctors. To alleviate this situation, the Deferred Acceptance Algorithm is used for doctor-patient matching. The algorithm takes into account the real preferences of both the doctor and the patient to ensure Strategy-Proofness and Pareto efficiency in the matching process. The patient applies to the doctor based on his or her preferences, and the doctor filters the patients based on his/her priorities, accepting the top-ranked patients for now and rejecting those who exceed the capacity. Rejected patients continue to apply to the next preferred physician until the matching process stabilizes.

Strategy-Proofness is a key attribute of Deferred Acceptance Algorithm, which ensures that participants honestly report the true preference as the optimal strategy, reducing the possibility of system manipulation. In healthcare resource allocation, strategy-independence reduces patients' incentives to exaggerate their conditions, optimizes resource utilization, and improves the efficiency and quality of overall healthcare services.

Pareto efficiency in the matching process ensures that no rematch can make some doctors and patients better off without harming others. This efficiency is achieved through the Deferred Acceptance Algorithm, which ensures fair matching results and efficient resource utilization, enhances trust between patients and physicians, and improves the stability and efficiency of the overall healthcare delivery system. Deferred Acceptance Algorithm can establish stable matching between patients and physicians despite the possibility of patients deceiving their preferences. However, patient behavior may affect the efficiency and fairness of the matching process, highlighting the importance of transparency and integrity of the doctor-patient matching system.

When it comes to resource allocation and solving the problems facing China's current healthcare system, Deferred Acceptance Algorithm is, in fact, an effective solution. In contrast, China's healthcare delivery system is markedly different from that of the U.K. Under the study by Ruth Leibowitz, Susan Day, and David Dunt, it was found that the RCT designed by the U.K. GP system, which integrates a nurse telephone counseling service (with the help of experienced, specially trained nurses using decision-support software) within an integrated healthcare co-op with the co-op's routine practice (receptionists recording call details and then passing them on to doctors) [10]. In other words, the allocation of doctor-patient resources is made by the GP, who first diagnoses and treats the patient and then refers the patient to a specialist. This allocation effectively reduces the tendency of patients to disguise their condition and helps to improve the efficiency of the whole healthcare delivery system. This is because patients cannot choose a specialist directly and must be referred through a GP. Even if a patient exaggerates his or her condition, he or she still needs to go through the assessment and judgment of the GP, whose professional judgment can, to a large extent, filter out unnecessary exaggeration and misreporting. This system ensures rational utilization of medical resources and improves the efficiency and fairness of overall medical services. At the same time, the key role of GPs in primary care and referral also enhances patients' trust in the healthcare system and reduces unnecessary doctor-patient conflicts. Therefore, the GP healthcare system in the UK is taken as the object of study in the following research. This will help to gain a deeper understanding of how the doctor-patient resource allocation mechanism works and can also provide references and insights for solving the problems faced by China's healthcare system.

References

[1] Han, Y., Lie, R.K., Li, Z. and Guo, R. (2022). Trust in the Doctor–Patient Relationship in Chinese Public Hospitals: Evidence for Hope. Patient Preference and Adherence, Volume 16, pp.647–657. doi:https://doi.org/10.2147/ppa.s352636.

[2] Futrell, G. D. (2021). Why Patients Lie: A Self-Discrepancy Theory Perspective of Patient Deception. Journal of Multidisciplinary Research, 13(2), 43-57. https://jmrpublication.org/wp-content/uploads/JMR13Fall2021002.pdf#page=45

[3] Jiang, Y.-P. and Yuan, D.-N. (2020). Surgeon-patient matching based on pairwise comparisons information for elective surgery. Computers & Industrial Engineering, 145, p.106438. doi:https://doi.org/10.1016/j.cie.2020.106438.

[4] Wong, T.C., Xu, M. and Chin, K.S. (2014). A two-stage heuristic approach for nurse scheduling problem: A case study in an emergency department. Computers & Operations Research, 51, pp.99–110. doi:https://doi.org/10.1016/j.cor.2014.05.018.

[5] Schummer, J. (1996). Strategy-proofness versus efficiency on restricted domains of exchange economies. Social Choice and Welfare, 14(1), pp.47–56. doi:https://doi.org/10.1007/s003550050050.

[6] Pathak, P.A. and Sönmez, T. (2008). Leveling the Playing Field: Sincere and Sophisticated Players in the Boston Mechanism. American Economic Review, 98(4), pp.1636–1652. doi:https://doi.org/10.1257/aer.98.4.1636.

[7] Robbins, L. (2007). An Essay on the Nature and Significance of Economic Science. [online] Google Books. Ludwig von Mises Institute. Available at: https://books.google.co.uk/books?hl=en&lr=&id=nySoIkOgWQ4C&oi=fnd&pg=PA1&ots=bzw3yZdpcx&sig=BLzCZcholle9ZnqwockSBGRtsn4&redir_esc=y#v=onepage&q&f=false [Accessed 17 Jun. 2024].

[8] Bó, I. and Hakimov, R. (2022). The iterative deferred acceptance mechanism. Games and Economic Behavior, 135, pp.411–433. doi:https://doi.org/10.1016/j.geb.2022.07.001.

[9] Balinski, M. and Sönmez, T. (1999). A Tale of Two Mechanisms: Student Placement. Journal of Economic Theory, 84(1), pp.73–94. doi:https://doi.org/10.1006/jeth.1998.2469.

[10] Ruth Leibowitz, Susan Day, David Dunt, A systematic review of the effect of different models of after-hours primary medical care services on clinical outcome, medical workload, and patient and GP satisfaction, Family Practice, Volume 20, Issue 3, June 2003, Pages 311–317, https://doi.org/10.1093/fampra/cmg313

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241562

Predicting financial enterprise stocks and economic data trends using machine learning time series analysis

Haotian Zheng[1,6], Jiang Wu[2], Runze Song[3], Lingfeng Guo[4], Zeqiu Xu[5]

[1]Electrical & Computer Engineering, New York University, New York, NY, USA
[2]Computer Science, University of Southern California, Los Angeles, CA, USA
[3]Information System & Technology Data Analytics, California State University, CA, USA
[4]Business Analytics, Trine University, AZ, USA
[5]Information Networking, Carnegie Mellon University, PA, USA

[6]hz2687@nyu.edu

Abstract. This paper explores the application of machine learning in financial time series analysis, focusing on predicting trends in financial enterprise stocks and economic data. It begins by distinguishing stocks from stocks and elucidates risk management strategies in the stock market. Traditional statistical methods such as ARIMA and exponential smoothing are discussed in terms of their advantages and limitations in economic forecasting. Subsequently, the effectiveness of machine learning techniques, particularly LSTM and CNN-BiLSTM hybrid models, in financial market prediction is detailed, highlighting their capability to capture nonlinear patterns in dynamic markets. Finally, the paper outlines prospects for machine learning in financial forecasting, laying a theoretical foundation and methodological framework for achieving more precise and reliable economic predictions.

Keywords: Machine learning, Financial time series analysis, LSTM, CNN-BiLSTM hybrid models, Stock market prediction

1. Introduction

Stocks represent ownership stakes in corporations issued to raise capital, entitling holders to residual profits and assets after debt obligations are fulfilled, alongside voting rights proportional to their shareholdings. Unlike stocks with fixed maturity dates, stocks do not expire as long as the issuing company remains solvent. Stocks guarantee fixed returns specified in the contract, whereas stock returns are variable and contingent upon corporate profitability and asset value [1]. This distinction underscores the inherent risk associated with stocks, where investors face uncertainty regarding dividends and capital gains contingent upon the company's financial performance. In contrast to risk-free stocks, which promise predictable returns albeit with default risk, there are no risk-free stocks. Stockholders assume greater risk due to fluctuating dividends and market prices, contingent upon broader economic conditions and company-specific factors.

Through empirical analysis and model validation, this research explores the effectiveness of machine learning techniques in capturing the nonlinear and dynamic nature of financial markets. Insights gained from this study contribute to academic discourse and have practical implications for investors, financial

DOI: 10.54254/2755-2721/87/20241562

analysts, and policymakers seeking to navigate volatile market conditions and optimize investment strategies. By harnessing AI's predictive capabilities, this paper aims to advance the understanding and application of machine learning in financial forecasting, paving the way for more accurate and reliable predictions in real-world economic scenarios.

2. Related work

2.1. Traditional Time Series Analysis Method

Traditional statistical methods have long been employed to forecast financial stocks and economic data because of their interpretability and historical reliability. Techniques such as the Autoregressive Integrated Moving Average (ARIMA) [2] model and exponential smoothing methods are notable examples. ARIMA models are adept at capturing linear trends and seasonality in time series data, making them suitable for predicting economic indicators with clear patterns over time.

Exponential smoothing methods, on the other hand, are effective in capturing short-term fluctuations and smoothing out noise in time series data. [3]They excel in scenarios where recent observations are more critical for forecasting than distant historical data points. Despite their strengths, these methods can be limited by their assumption of stationary data and may not adequately handle complex, non-stationary economic data.

Advantages and Limitations:

The strengths of traditional statistical methods lie in their interpretability and well-established theoretical foundations. They provide insights into economic variables' underlying trends and patterns, aiding decision-making processes. [4]Trendlines help illustrate long-term patterns, while seasonal decomposition charts break down the time series into trend, seasonal, and residual components, aiding in identifying cyclical patterns.

In summary, while traditional statistical methods have been foundational in economic forecasting, their limitations in handling non-linear and dynamic relationships have prompted the exploration of more advanced techniques, including machine learning models. The following sections will delve into how machine learning, particularly multivariate time series models, addresses these challenges and offers new opportunities for improving the accuracy and robustness of economic and financial predictions.

2.2. Application of machine learning to financial forecasting

Time series model: For multidimensional time series data, time series models in machine learning such as ARIMA, LSTM, GRU, etc. can be used for modeling and prediction. These models capture temporal relationships between data to make predictions about future trends.

(1) LSTM (Long Short-Term Memory)

LSTM is a special type of recurrent neural network (RNN) [5]. Unlike traditional neural networks, RNNS have a cyclic structure that allows information to flow continuously through the network. This allows RNNS to retain information over a long period theoretically. In practice, however, standard RNNS have trouble capturing long-term dependencies. LSTM was designed to solve this problem.

LSTM architecture:

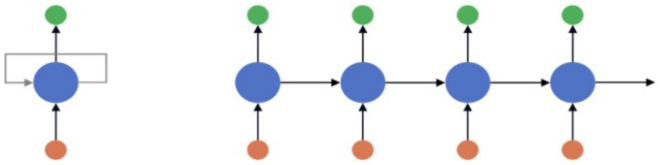

Figure 1. The basic network architecture of RNN

The LSTM architecture consists of a single unit, the memory unit (also known as the LSTM unit). The LSTM unit includes four feedforward neural networks. Each network has input and output layers, with connections from input neurons to all output neurons.

DOI: 10.54254/2755-2721/87/20241562

In predicting financial market fluctuations and trends, the effect of LSTM is mainly reflected in its ability to deal with complex non-linear relationships and dynamic market environments [6]. Compared to traditional linear models or simple time series methods, LSTM can better adapt to changing patterns and behaviors in financial markets. Its ability lies in efficiently using historical data and external influences through multi-level feature learning and memory mechanisms to provide more accurate and dynamic prediction results [7]. This capability makes LSTM a powerful tool in financial forecasting, helping investors and analysts better understand market dynamics and make more informed investment decisions.

(2) Neural network(CNN)

The neural network model is a powerful machine learning model that can effectively deal with complex nonlinear relationships and large-scale data. Through multi-level neuronal structures and nonlinear activation functions, neural networks can learn and express complex patterns and higher-order features in data to adapt to various data distributions and complex relationships.

In addition, the training process of neural networks can be time-consuming, especially when dealing with large data sets that require high-performance computing resources. In the financial market prediction, applying a neural network model also needs to consider the market's complexity and the data's uncertainty, as well as the generalization ability and stability of the model [8]. In summary, neural network models provide new perspectives and tools for financial market prediction through their powerful nonlinear modeling capabilities, data-driven methods, and more accurate and comprehensive analytical support for investment decision-making and risk management.

2.3. Time series and financial forecasting

With the continuous development of machine learning technology, the application of machine learning in the financial field has also made remarkable progress. These applications improve the prediction accuracy and simplify the application process of traditional models. In this paper, we will list some effective applications of machine learning technology in the financial field from the aspects of factor extraction, missing value filling, fusion input, noise reduction, and non-independent co-distributed adaptation to help readers better understand the current situation of machine learning in the financial field [9]. In addition, tensor-filling methods have been applied in finance to fill in patio-temporal data and show obvious potential. Although some efforts have been made in this area, research is still relatively limited. Advanced deep learning methods can introduce nonlinear and patio-temporal interactions to fill in missing values in financial data.

2.4. Stock forecasting-related tasks

Before delving into the details of deep learning models, we will first define four key stock market prediction tasks and outline the concepts associated with each task [10]. These tasks include stock price forecasting, stock trend forecasting, portfolio management, and trading strategies, and these categories summarise most existing stock market forecasting tasks.

- Stock price forecasting uses time series data to predict the future value of stocks and financial assets traded on exchanges. The goal of this forecast is to achieve a healthy profit. In addition, various factors also affect the forecasting process, including psychological factors and rational and irrational behavior, all of which work together to make stock prices dynamic and volatile.
- Forecasting stock movements typically divides stock trends into three categories: uptrend, downtrend, and sideways; this task is formalized by analyzing the difference between adjusted stock closing prices within a given trading day.

In the stock market prediction task using deep learning, common trading strategies include event-driven, data-driven, and strategy optimization, and the above tasks revolve around the stock market prediction process. The next step is to input extracted features into the deep learning model for training, and finally, the experimental results of the model are analyzed.

DOI: 10.54254/2755-2721/87/20241562

3. Methodology

In recent years, applying advanced machine learning techniques to financial time series analysis has garnered significant attention due to their potential to uncover intricate patterns and improve prediction accuracy. Among these techniques, the combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs), known as CNN-LSTM models, has proven particularly effective. This approach leverages CNNs to extract spatial features from input data and LSTMs to capture temporal dependencies. It is well-suited for analyzing multivariate time series data such as stock prices and economics.

3.1. Model discussion

The CNN- BiLSTM LSTM model integrates two robust neural network architectures:

1. Convolutional Neural Networks (CNNs):

- CNNs are adept at learning spatial hierarchies of features through convolutional layers.
- In the context of multivariate time series, CNNs can be applied to extract spatial patterns across different variables (e.g., multiple stock prices, economic indicators) at each time step.

2. Long Short-Term Memory networks (LSTMs):

- LSTMs are well-suited for modeling temporal dependencies by maintaining long-term memory of sequential data.

3. BiLSTM: Based on the cell structure of LSTM, the LSTM historical model has more vital historical information screening ability and chronological order learning ability and can rationally use the input historical data information to form long-term memory of historical data information in the past period, thus avoiding the problem that effective historical information cannot be stored permanently due to the influence of continuous input historical data. Since data processing depends on the direction of network connection, Bi-directional Long Short-Term Memory (BiLSTM) is introduced for events that need to consider the impact of future data on historical data. The model can reference the influence of both historical and future data on the predicted results.

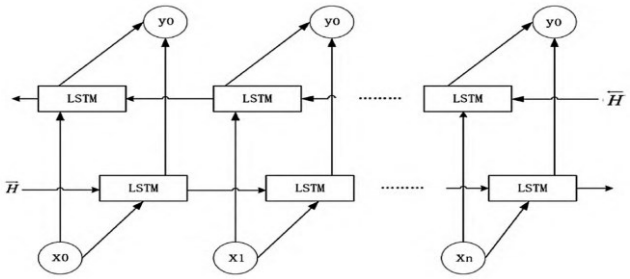

Figure 3. BiLSTM expansion structure based on LSTM

The data is preprocessed first, including the processing of missing and duplicate values, and then normalized. After the processing, the data is divided into a test set and a training set. The attention mechanism is used to increase the weight value of the extracted features and update the weight. Finally, the CNN-BiLSTM-Attention model was constructed, and the test set data was input to verify the model's accuracy.

3.2. Data processing

In this experiment, the stock data of the People's Bank of China were studied. The research results were compared with the LSTM model (LSTM-ATTENTION), the convolutional neural network, and the bidirectional long and short memory neural network mixed model (CNN-BiLSTM), and the single long and short memory neural network model (LSTM). It is concluded that the CNN-BiLSTM-Attention model has a good effect.

Due to the significant difference in the results of the stock data, the data needs to be normalized before input into the neural network model using 0-1 normalization. The calculation method is as follows:

DOI: 10.54254/2755-2721/87/20241562

$$x^x = \frac{x - min}{max - min} \tag{1}$$

x is the original sample data value; min is the minimum value in the sample data.max indicates the maximum value in the sample data.

3.3. Test data and methods

The input of the neural network is the data closely related to the trading of the stock, and the output is the closing price, which predicts the close of the next trading day's price. This paper downloads experimental data from Tushara's official website. It selects the data of People's Bank of America Stock (stock code 000001) from January 1, 2005, to October 4, 2021, of which 80% is used as the training set and 20% is used as the test set.

In this paper, the window size is n, and roll is selected. The window size is 1, and MAE comparisons of different window lengths are selected 5 times. See Table 1.

Table 1. Comparison of results of different sliding window sizes

Sliding Window Size	Mean Absolute Error (MAE)
5	0.01945795
7	0.01809513
10	0.01776377
15	0.02096804

Table 1 shows that when the window size is 5, the MAE value is larger than the step size. When the selection is 15, the MAE value is also relatively large, and when the selection step is 10, the average absolute error value is the smallest, so the optimal window size is selected Select 10.

To verify the high accuracy of the model, different algorithms are used for comparison, and the comparison results are shown in Table 2.

Table 2. Comparison of predictions from different models

Model	MSE (Mean Squared Error)	MAPE (Mean Absolute Percentage Error)
CNN-BiLSTM-Attention	0.012864103	0.0198415
LSTM-Attention	0.03095998	0.0242836
CNN-BiLSTM	0.071255782	0.059437
LSTM	0.031070495	0.0816299

As can be seen from Table 2, compared with the LSTM hybrid model, the overall trend is better, while the new hybrid model CNNBiLSTM-Attention model MSE is 0.012864103, MAPE is 0.01984150. It has higher reliability than previous models.

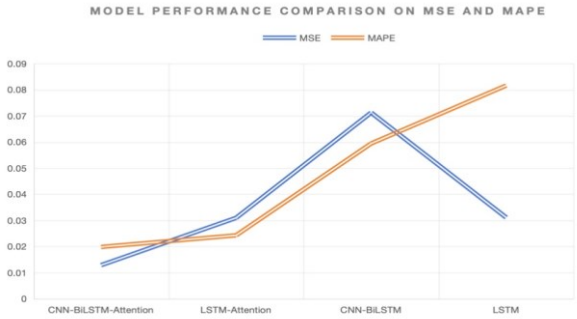

Figure 4. Evaluation of Model Accuracy: MSE vs MAPE

DOI: 10.54254/2755-2721/87/20241562

3.4. Experimental design

By comparing the performance of the attention-based convolutional neural network and bidirectional long and short-memory neural network mixed models with traditional statistical methods, we draw the following conclusions:

- Improved prediction accuracy: Experimental results show that the new hybrid model significantly improves the accuracy of predicting changes in stock prices. Compared with traditional statistical methods, the model performs better on several evaluation indicators, such as mean square and absolute percentage errors.
- Effectiveness of feature extraction: Using convolutional neural networks (CNNs) for feature extraction can effectively capture spatial information in the input data, which is particularly important for analyzing multivariate time series. These extracted features help improve the subsequent model's predictive power (BiLSTM).
- Timing modeling of BiLSTM models: Bidirectional Long and short memory neural networks (BiLSTM) perform well in processing time series data, effectively capturing long and short-term timing dependencies, thereby improving the robustness and accuracy of predictions.

The hybrid CNN-BiLSTM-Attention model demonstrates clear advantages in processing financial time series data and predicting quantitative transactions.

4. Conclusion

Based on the considerations for long-term stability and reliable forecasting in stock markets, it is evident that short-term stock price predictions can inadvertently promote short-sighted investor behavior. This tendency undermines the market's long-term stability and hampers its sustainable growth. Developing robust long-term forecasting models that incorporate multiple influencing factors is crucial to counteract this. These models should transcend the immediate fluctuations and provide insights contributing to a more stable and predictable market environment.

Furthermore, as China's capital markets continue to undergo reforms and development, a significant imperative remains to refine market institutions through ongoing exploration. Achieving high-quality development in China's distinctive modern capital market requires a sustained commitment to these principles, ensuring a balanced approach that supports long-term investor confidence and economic resilience.

References

[1] Colladon, Andrea Fronzetti, and Giacomo Scettri. "Look inside. Predicting stock prices by analyzing an enterprise intranet social network and using word co-occurrence networks." International Journal of Entrepreneurship and Small Business36.4 (2019): 378-391.

[2] Ariyo, Adebiyi A., Adewumi O. Adewumi, and Charles K. Ayo. "Stock price prediction using the ARIMA model." 2014 UKSim-AMSS 16th international conference on computer modeling and simulation. IEEE, 2014.

[3] Khare, Kaustubh, et al. "Short-term stock price prediction using deep learning." 2017 2nd IEEE international conference on recent trends in electronics, information & communication technology (RTEICT). IEEE, 2017.

[4] Qian, K., Fan, C., Li, Z., Zhou, H., & Ding, W. (2024). Implementation of Artificial Intelligence in Investment Decision-making in the Chinese A-share Market. Journal of Economic Theory and Business Management, 1(2), 36-42.

[5] Xu, Jiahao, et al. "AI-BASED RISK PREDICTION AND MONITORING IN FINANCIAL FUTURES AND SECURITIES MARKETS." The 13th International scientific and practical conference "Information and innovative technologies in the development of society"(April 02–05, 2024) Athens, Greece. International Science Group. 2024. 321 p.. 2024.

[6] Wang, Yong, et al. "Machine Learning-Based Facial Recognition for Financial Fraud Prevention." Journal of Computer Technology and Applied Mathematics 1.1 (2024): 77-84.

DOI: 10.54254/2755-2721/87/20241562

[7] Song, Jintong, et al. “LSTM-Based Deep Learning Model for Financial Market Stock Price Prediction.” Journal of Economic Theory and Business Management 1.2 (2024): 43-50.
[8] Shi, C., Liang, P., Wu, Y., Zhan, T., & Jin, Z. (2024). Maximizing user experience with LLMOps-driven personalized recommendation systems. Applied and Computational Engineering, 64, 101-107.
[9] Tian, J.; Li, H.; Qi, Y.; Wang, X.; Feng, Y. Intelligent medical detection and diagnosis assisted by deep learning. Appl. Comput. Eng. 2024, 64, 116–121, https://doi.org/10.54254/2755-2721/64/20241356.
[10] Tavenard, R., Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., ... & Woods, E. (2020). To learn, a machine learning toolkit for time series data. Journal of Machine Learning Research, 21(118), 1-6.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241543

Decision tree C4.5 algorithm for generative AI technology ethics -- Based on the results of the questionnaire

Sihan Xu

North China University of Technology

806484259@qq.com

Abstract. With the development of AI technology, generative AI has gradually entered the life of the public, for example, the explosion of CHAT-GPT has allowed more people to see the huge potential and obvious advantages of generative AI. However, in the process of generative AI operation, events that violate social responsibility and ethics often occur, which makes the research on the scientific and technological ethics of generative AI more urgent. In the past literature and research, many industry experts have analysed the impact of generative AI on specific industries, but everyone is or will be a user of generative AI, so we should pay attention to the study of the people's scientific and technological ethical issues of generative AI after putting aside the industry background, so this paper collects primary data by means of questionnaire surveys to find out the public's awareness of generative AI and their perception of generative AI. and attitudes towards generative AI, and using the decision tree C4.5 algorithm with Python as the tool, it is used to respond to people's awareness of generative AI and the public's perception of the relationship between the various factors of the ethical issues of

Keywords: Generative AI, Decision Tree Algorithms, Ethics of Technology

1. Introduction

Back in 2018, Ming-Hui Huang [1] discussed the impact of AI's tech ethics in the service industry in an article examining the substitutability of AI in terms of machines, foreseeing that in the future in the service industry, some simple tasks will be taken over by AI, resulting in the loss of personnel, which is seen as a transitional phase of augmentation, and then, when it has the ability to take over all the work tasks it will then completely replace human labour.

By 2022, Martin Reisenbichler [2] shows in his research that the emergence of AI has enabled the realisation of natural language generation to support content marketing, a study of the ethics of AI from the field of writing, pointing out that although machine-generated content is designed to perform well in search engines, the role of human editors is still vital.

Next in 2024, targeting the aspect of literary creation, Holden Thorp [3] tried in his experiment to ask generative AI to rewrite the first scene of the classic American play Death of a Salesman, but with Princess Elsa from the animated film Frozen as the main character instead of Willy Loman, which resulted in an amusing dialogue in which Elsa returns home after a hard day of selling, and her son, Harpy, says to her, "Come on, Mum, you're Elsa from Frozen. You have the power of ice and snow and you are the queen. You're unstoppable." Mashups like this are certainly interesting, but they have serious implications for generative AI programs like ChatGPT in science and academia. And not just in terms

DOI: 10.54254/2755-2721/87/20241543

of literary creation. In the same year, also in terms of literary creation, Chris Stokel-Walker [4] noted that as generative AIs like ChatGPT become popular in 2023, major ethical discussions about their role in academic authorship have emerged. Prominent ethical organisations, including the ICMJE and COPE, as well as leading publishers, have developed ethical clauses that make it clear that these models don't meet authorship standards due to accountability issues.

Next, in 2024, Shiavax J Rao [5] argues in his article that AI, and in particular high-level language models like ChatGPT, have the potential to revolutionise all aspects of healthcare, medical education and research, reviewing the benefits of ChatGPT in personalising patient care, particularly in geriatric care, medication management, weight loss and nutrition, and sports activity instruction, with further insights into its potential for enhancing medical research through the analysis of large datasets and the development of new methods. In the field of medical education, to make ChatGPT an effective resource for medical students and professionals as an information retrieval tool and for personalised learning.ChatGPT has many promising applications that may trigger a paradigm shift in healthcare practice, education and research.The use of ChatGPT may be beneficial in the areas of clinical decision-making, geriatric care, medication management, weight loss and nutrition, physical fitness scientific research, and medical education, among other areas. However, it is worth noting that issues around ethics, data privacy, transparency, inaccuracy and inadequacy remain. The real-world impact of ChatGPT and generative AI must be objectively assessed using a risk-based approach before it can be widely used in medicine. It is not difficult to see that AI ethical issues, gradually from the ethical issues of man and machine, to the field of literary creation and then to the medical and other fields with greater relevance to people, generative AI technology ethical issues are destined to become an unavoidable problem for people in the future.

2. Research design

2.1. Analysis of questionnaire results

2.1.1. Questionnaire data collation

The data for this study came from a questionnaire that categorised the main reasons affecting the ethics of science and technology into five categories in the form of scales: "access to discriminatory search results", "inaccurate information answered", "misuse of information", "false content and malicious dissemination", and "impact on a person's ability to make autonomous decisions", with five specific measurable indicators in each category. misuse of information", "false content and malicious dissemination", and "impact on people's ability to make autonomous decisions", which are five specific and measurable indicators, and the sub-questions in each category are measured by a score from 1 to 10, and the average value will be calculated. degree, and derive the average value, classifying 1~4 as mild (a), 5~7 as moderate (b), and 8~10 as severe (c), and at the same time, respectively, using A=obtaining discriminatory search results, B=inaccurate information in the answers, C=information misuse, D=false content and malicious dissemination, and E=influence on people's autonomous decision-making ability, and, the questionnaire's "Perceived importance of AI" column, with scores of 1 to 5 as unimportant and scores of 6 to 10 as important, to facilitate the next decision tree C4.5 algorithm.

2.1.2. Questionnaire design and distribution

A total of 270 questionnaires were distributed in this research study, with response time of more than 200 seconds as the detection criterion, and excluding the text papers that did not meet the criteria, leaving a total of 221 valid questionnaires. The questionnaires were divided into three levels, namely young (10-28 years old), middle-aged (29-47 years old), and old (48-66 years old), and 30% of the respondents from each age group were selected as the final database, of which 89 were selected from the young, 110 from the middle-aged, and 22 from the old, and the questionnaires were taken as non-scaled questions to collect information and data analysis was done using SPSS.

DOI: 10.54254/2755-2721/87/20241543

2.2. Data analysis

2.2.1. Summary of basic information

Table 1. Analysis of Respondents by Industry

Industry Name	frequency	Percentage (%)
Internet technology industry	32	14.5
financial industry	56	25.3
Consultancy services industry	41	18.6
Education Industry	46	20.8
Government and public interest organisations	23	10.4
student at school	17	7.7
the rest	6	2.7
(grand) total	221	100

In terms of industry distribution, 221 questionnaires were sent and 221 were valid, with the financial industry having the most practitioners with 56; the education industry followed with 46; and the third was the counselling services industry with 41, as shown in Table 2.

2.2.2. Next, let's look at the percentage of people who have used generative AI in each industry:

Table 2. Status of use of generative AI by industry

	Internet technology industry		financial industry		Consultancy services industry	
	frequency	per cent	frequency	per cent	frequency	per cent
used up	20	10	30	13.6	20	9
unused	10	4.5	26	11.8	21	9.5
	Education Industry		Government and public interest organisations		student population	
	frequency	per cent	frequency	per cent	frequency	per cent
used up	19	8.6	11	5	7	3.2
unused	27	12.2	12	5.4	9	4.1

Through the above table, we can see that although the number of employees in the financial industry is large, the number of people who have used generative AI and those who have not used it each accounts for about half, but the percentage of those who have used generative AI in the Internet industry reaches 10%, which is 5.5% higher than that of those who haven't used it, and the difference between the percentage of those who have used it and the percentage of those who haven't is -0.5%, -3.6%, and -0.4% respectively, indicating that the most widely exposed to and using generative AI is the group in the Internet technology industry. The percentage difference between those who have used it and those who haven't is -0.5%, -3.6%, and -0.4% respectively, indicating that the most widely exposed to and used generative AI is the group in the Internet technology industry.

DOI: 10.54254/2755-2721/87/20241543

Table 3. Statistics on the number of people who expect to use generative AI again

	frequency	Effective percentage
lesser	43	38.7
usual	40	36
non-recurrent	26	23.4
I can't get away from it.	2	1.8
(grand) total	111	100

The table shows that generative AI is not used very often among the people surveyed, and the number of people who use it generally versus less often amounts to 83, or 74.7%.

How many of those using generative AI have heard of the concept of tech ethics, as shown in Table 4.

Table 4. Perceptions of technology ethics among those who have used generative AI

	frequency	Effective percentage
be	55	50
clogged	55	50
(grand)total	110	100

As can be seen from the table above, there is exactly a 50/50 split between those who have heard of tech ethics and those who have not, but since those who have used generative AI accounted for 49.4% of the total survey respondents, it suggests that tech ethics is still a relatively new concept in the general public's perception.

2.3. Data modelling

We take the decision tree C4.5 model to measure the importance people attach to the ethical issue content of different generative AIs, so we have to calculate by INFO information, e information entropy, Gain information gain, Gain Rate information gain rate, and Gini coefficient [12].

First of all, we should calculate the binary classification result: whether it is important to organise the science and technology ethics of generative AI into "important" and "unimportant", which results in the number of important people as H, and the number of unimportant people as J. Then the total amount of information will be as in Expression 1.

$$INFO_{总} = I[H,J] = \frac{H}{H+J} log_2\left(\frac{H}{H+J}\right) - \frac{J}{H+J} log_2\left(\frac{J}{H+J}\right) \tag{1}$$

Next, we want to calculate the information entropy. Each dataset is converted into a rank 1\2\3 (0~4 is classified as "1", 5~8 is classified as "2", and 9~11 is classified as "3") based on the scores Assuming that the set E $\in$ ({A},{B},{C},{D},{E}), P_i is a probability distribution, the information entropy ENT (E) is as in Equation 2.

$$ENT(E) = \sum_{I=1}^{3} P_i \, log_2(P_i) \tag{2}$$

After finding the information entropy $ENT(E)$, we calculate its information gain as in Equation 3.

$$Gain(E) = INFO_{总} - ENT(E) \tag{3}$$

Next, the value of the split information is derived by setting the number of important and unimportant ratings corresponding to each classification to L_1 = number of important and L_2 = number of unimportant ratings, as in Equation 4.

DOI: 10.54254/2755-2721/87/20241543

$$SplitINFO_E = -\frac{L1}{L1+L2}log_2\left(\frac{L1}{L1+L2}\right) - \frac{L2}{L1+L2}log_2\left(\frac{L2}{L1+L2}\right) \quad (4)$$

Using the information on it, the information gain ratio is calculated as in Equation 5.

$$GainRate(E) = \frac{Gain(E)}{SplitINFO_E} \quad (5)$$

Meanwhile, the Gini coefficient is calculated as in Equation 6.

$$Ginicoefficient = 1 - \sum(pi\text{^}2) \quad (6)$$

2.4. Operating Mechanisms

In the above manner, repeat the calculation, select MAX as the node, to divide the root node, leaf nodes and the end point, until a branch is all "important" or "unimportant" so as to gradually draw a decision tree.

2.4.1. Data entry

Firstly, the data is divided into test set and training set in the ratio of 1:9. The parameters are designed as follows:

In the second step, in the training set, the root node is selected and the calculated values are given in Table 5 below.The set E has the largest gain information rate, so it is selected as the root node.

Next continue to repeat this arithmetic rule, you can plot the decision tree, as in Figure 1.

Table 5. Selection of root nodes

	ENT(E)	Gain(E)	SplitINFO$_E$	$GainRate$	Weight value
A	0.26	0.037	1.416	0.026	0.272
B	0.134	0.163	1.393	0.117	0.139
C	0.125	0.172	1.436	0.120	0.162
D	0.132	0.165	1.450	0.113	0.151
E	0.14	0.157	1.163	0.135	0.276

DOI: 10.54254/2755-2721/87/20241543

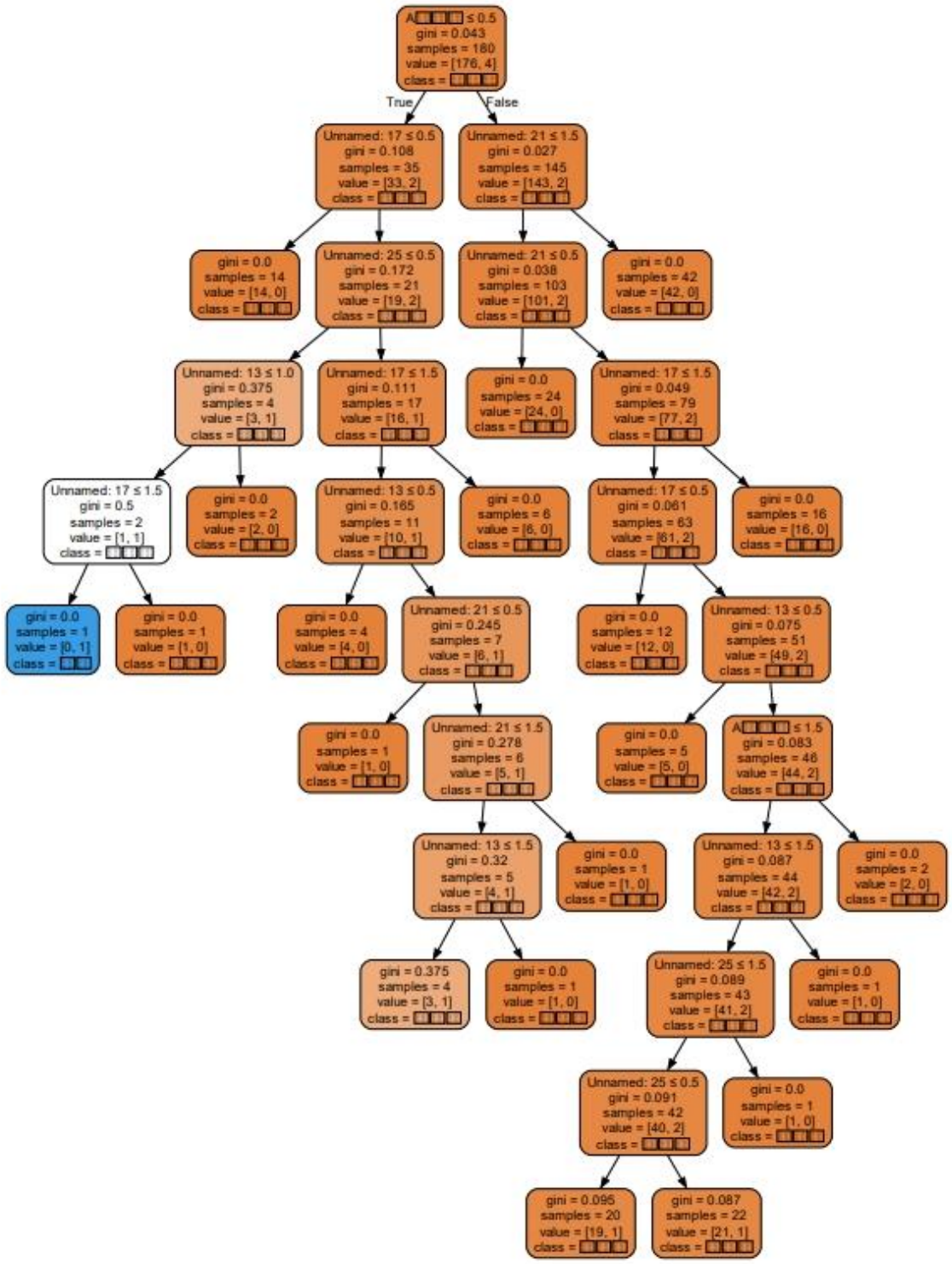

Figure 1. Decision Tree Model of Influencing Factors for Generative AI Tech Ethics

2.4.2. Prune (branches etc)

When using decision tree algorithms for classification or regression problems, pruning techniques are often employed to avoid overfitting. ccp_alpha is a pruning parameter that controls the strength of the pruning.

DOI: 10.54254/2755-2721/87/20241543

For each leaf node, we can compute an effective value with respect to ccp_alpha. Specifically, the effective value of each node is its reduced impurity (e.g., Gini impurity) minus a scaling factor, ccp_alpha, which is multiplied by the number of offspring of that node. Thus, if a node's effective value is less than zero, then that node will be pruned. the larger the ccp_alpha, the stronger the pruning and the simpler the final decision tree; conversely, the smaller the ccp_alpha, the weaker the pruning and the more complex the final decision tree.

If a node has a negative valid value, the node will be pruned out. Note that this is just an example and in practice different datasets and models may require different ccp_alpha values.

When using decision tree algorithms for classification or regression problems, we can control the complexity and generalisation ability of the model by adjusting the ccp_alpha parameter. Typically, we use techniques such as cross-validation to select the best ccp_alpha value to achieve the best model performance and generalisation ability.

2.4.3. Model testing

Table 6. Training set model evaluation results

term (in a mathematical formula)	accuracy	recall rate	f1-score	sample size
usual	0.61	0.59	0.60	37
significant	0.90	0.90	0.90	143
accuracy			0.84	180
average value	0.75	0.75	0.75	180
Average (combined)	0.84	0.84	0.84	180

Table 7. Test set model evaluation results

term (in a mathematical formula)	accuracy	recall rate	f1-score	sample size
usual	0.50	0.33	0.40	3
significant	0.89	0.94	0.92	18
accuracy			0.86	21
average value	0.70	0.64	0.66	21
Average (combined)	0.84	0.86	0.84	21

Table 8. Integrated model assessment

name (of a thing)	parameter name	parameter value
Model parameter setting	Data preprocessing	Norm
	Training set ratio	0.9
	Nodal split criteria	Gini
	Node division method	Best
	Minimum number of samples for node splitting	2
	Leaf node minimum sample tree	3
	Maximum tree depth	10
Modelling to assess effectiveness	accuracy	85.714 per cent
	Precision rate (combined)	83.835 per cent
	Recall rate (combined)	85.714 per cent
	f1-score	0.845

DOI: 10.54254/2755-2721/87/20241543

The above tables show the performance of the model on the training set and test set respectively.

First: accuracy rate, the proportion of samples with correct prediction results to the total samples, the accuracy rate training set is 0.84, the test set is 0.86, which belongs to the higher accuracy rate; Second: precision rate, the prediction results are positive in the results of the training set and the test set is greater than 0.5, the precision is better; Third: recall rate, the proportion of positive samples with positive predictions, except for the test set which is "general" is 0.33, the others are all greater than 0.5, indicating that the model has a high recall; Fourth: f1-score, is a comprehensive evaluation index that integrates the precision rate and the recall rate it is the reconciled average of the precision rate and the recall rate; Fifth: the higher the precision rate and the recall rate are, the better, but the two tend to contradict each other, so the f1-score is commonly used to integrate the precision rate and recall rate. Commonly used f1-score to comprehensively evaluate the effect of the classifier, which takes the value of the range of 0 to 1, the closer to 1 the better the effect, so the improvement of the model is better.

So in synthesis, it can be seen that: the final model obtained an accuracy of 85.71% on the test set, a precision (combined) of 83.83%, a recall (combined) of 85.71%, and an f1-score (combined) of 0.84. The model results are acceptable.

3. Conclusion

According to the model, among the people who think generative AI is important, item E "impact on human autonomous decision-making ability" is the most important impact of generative AI in technology ethics, and those who are concerned about item E are also more concerned about item A "obtaining discriminatory search results", followed by item C "misuse of information", while item B "inaccurate answer information" and item D "false content and malicious dissemination" are considered by people to be the most important impacts of generative AI. and "obtaining discriminatory search results" in item A, followed by "misuse of information" in item C. People are less concerned about "inaccurate information" in item B and "false content and malicious dissemination" in item D. Therefore, in practice, people will be more concerned about the "misuse of information" in item B and "inaccurate information" in item D. Therefore, in practice, if we want people to pay attention to the ethics of generative AI, we should focus on the impact of generative AI on human autonomy and the avoidance of discriminatory answers when formulating the ethical norms of generative AI, rather than just guaranteeing that the output of generative AI is accurate or not in order to govern the industry. At the same time, it is also necessary to strengthen people's attention to the misuse of information and privacy protection, such as: the establishment of science and technology ethics education courses in colleges and universities, included in the mandatory curriculum, and at the same time, the establishment of WeChat public number, or in the short video number of the continuous release of a series of science and technology ethics education courses and other ways to subconsciously influence the importance of science and technology ethics in the minds of the people.

References

[1] Huang, M.-H., & Rust, R. T. (2018). Artificial Intelligence in Service. Journal of Service Research, 21(2), 155-172.

[2] Aaker JL (1997) Dimensions of brand personality. J. Marketing Res. 34(3):347–356.

[3] Nazarovets, S., & Teixeira da Silva, J. A. (2024). ChatGPT as an “author”: Bibliometric analysis to assess the validity of authorship. Accountability in Research, 1–11.

[4] Stokel-Walker, C., & van Noorden, R. (2023). What ChatGPT and generative AI mean for science. Nature, 614, 214-216.

[5] Rao, S.J., Isath, A., Krishnan, P. et al.(2024) ChatGPT: A Conceptual Review of Applications and Utility in the Field of Medicine. J Med Syst 48-59.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241523

Communication security analysis of fully electronic interlocking systems

Xiyan Hou

Key Laboratory of Photoelectric Technology and Intelligent Control of the Ministry of Education, Lanzhou Jiaotong University, Lanzhou, China

2073787258@qq.com

Abstract. Communication security in fully electronic interlocking systems is one of the key factors ensuring the safe and reliable operation of the entire system. EN50159 is an important standard in the European railway communication sector, aimed at ensuring the safety, reliability, and efficiency of railway transportation. According to the communication security standards of EN50159, there are six types of security risks in closed communication: data duplication, deletion, insertion, misordering, delay, and corruption. This paper analyzes and explains the aspects of communication security that need to be considered based on the safety communication standards of China's railway signaling system and the signal safety standards in EN50159, focusing primarily on communication security and reliability.

Keywords: Fully electronic interlocking system, EN50159, closed communication, communication security

1. Introduction

With the rapid development of computer networks and communication technology, railway signaling systems have greatly improved, making train operations faster, safer, and more efficient. The advent of new technologies in modern communications and microelectronics has accelerated the development of computer networks and communication technology, driving continuous upgrades in railway signaling technology. Communication-based train control systems have seen broader application, though the relationships between railway signaling systems have become more complex [1]. The adoption of advanced fully electronic computer interlocking systems, which no longer rely on traditional gravity-based safety relays but use electronic execution units for ultimate control, offers significant advantages in terms of maintainability, reduction of control room area, and construction workload. These systems have become the mainstream direction for railway signal control systems in China [2]. Given their high safety and real-time requirements, the operating cycle of fully electronic computer interlocking systems must be less than 250ms, and the safety performance must meet the SIL4 standard [3].

2. Safety Communication Standards

Railway communication systems are primarily used for train control, signal transmission, personnel communication, and emergency rescue. These systems must be highly reliable, stable, and resistant to interference to ensure the safety and smooth operation of railway transport. With the establishment of safety standards, international organizations have developed various versions of safety communication

DOI: 10.54254/2755-2721/87/20241523

protocols tailored to different train control systems, transmission networks, and defense against attacks, ensuring communication safety in railways. EN50159 is a crucial standard in the European railway communication field, aimed at ensuring the safety, reliability, and efficiency of railway transportation. It provides a comprehensive safety assurance system for railway information transmission systems with its strict functional requirements and technical specifications [4]. The EN50159 standard includes requirements for the design, installation, operation, and maintenance of communication systems, providing a unified reference framework to ensure compatibility and interoperability among different systems within the railway industry. This standard covers various communication technologies, including wired and wireless communication systems, and equipment related to train control, signal transmission, and personnel communication. To ensure secure communication transmission, we must strictly adhere to security communication protocols and ensure the stability of the characteristics of the closed transmission system, so that the number of connected devices and the maximum data capacity are not affected [5], thereby reducing the risk of illegal interference.

To ensure the safety of closed transmission systems, we must detect and prevent risks such as data frame overlap, omission, insertion, confusion, error, and timeout as early as possible. These risks include, but are not limited to, transmission system failures and external influences. Therefore, before designing communication protocols, it is essential to carefully review the characteristics of data frames for accuracy, reliability, orderliness, and timeliness to ensure the entire communication system meets safety communication requirements [6].

3. Fully Electronic Computer Interlocking System

3.1. Fully Electronic Computer Interlocking System Architecture

In traditional railway signaling systems, train operations are manually controlled by signal operators. In contrast, fully electronic computer interlocking systems achieve signal control and train dispatch through electronic devices and computer software. These systems have significant advantages in improving operational efficiency, reducing human error, and enhancing safety [7]. The fully electronic computer interlocking system typically consists of the following main components [8]:

1. **Computer System:** Responsible for controlling and managing the entire interlocking system, including functions such as processing train location information and signal control commands.
2. **Interface Equipment:** Communicates with track equipment, signal devices, and train location detection devices to obtain real-time train location and status information.
3. **Interlocking Logic Control Unit:** Formulates signal control logic based on train location, dispatch plans, and other information to ensure safe and smooth train operations.
4. **Communication Equipment:** Facilitates communication among various parts of the system, including data transmission and command delivery.
5. **Human-Machine Interface:** Provides an interface for operators to monitor and manage the system. Typically, it displays train locations, signal status, and other information on a computer screen, allowing operators to take appropriate actions based on system prompts.

In summary, the structure of the fully electronic computer interlocking system is shown in Figure 1. The fully electronic communication layer communicates with the electronic control module layer via a bus, and both the interlocking logic layer and the fully electronic communication layer adopt a 2x2 redundancy structure. The electronic control module layer uses a dual-machine hot standby structure, with both the interlocking machine and the communication machine employing safety computers [9].

DOI: 10.54254/2755-2721/87/20241523

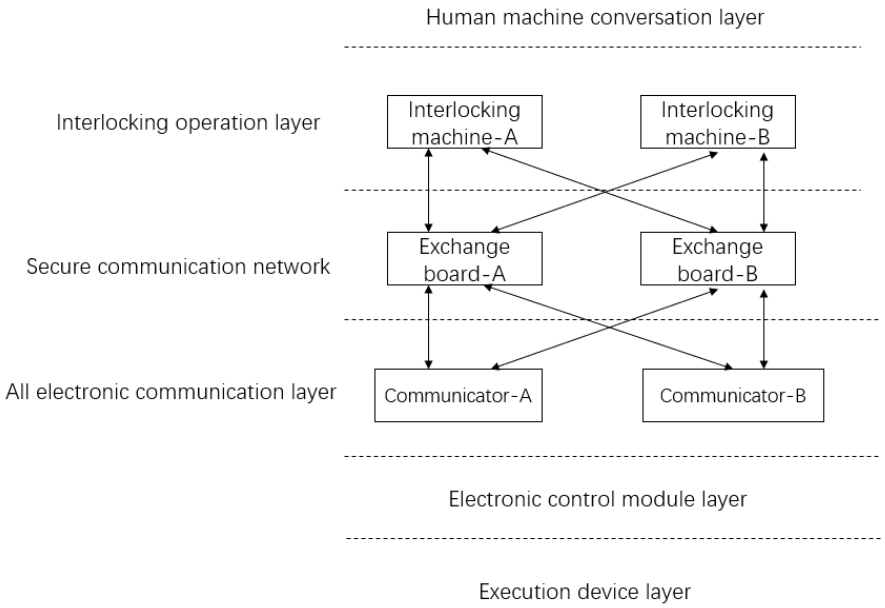

Figure 1. Structure of the Fully Electronic Computer Interlocking System

3.2. Safety Communication Requirements of the Fully Electronic Computer Interlocking System

The safety communication network of the fully electronic interlocking system is designed for communication between the interlocking machine and the communication machine, forming a closed transmission network [10]. It follows the EN50159 technical standard set by the European Committee for Electrotechnical Standardization (CENELEC) to ensure safety at railway crossings. EN50159 is an important standard in the European railway communication signaling field, outlining the basic requirements for safe communication protocols to ensure system safety and reliability. Currently, some European equipment or system solutions used in China's train control systems involve the safety communication system and interface protocol established by the EN50159 standard. This standard not only clearly identifies the potential dangers of closed transmission systems but also determines the best protective measures based on this technical standard to ensure railway operational safety [10].

According to the EN50159 standard, in a closed communication environment, to reduce threat risks, safety function modules are generally embedded in the application layer and the communication protocol data layer to implement safety protocols. The safety function modules can provide four types of verifications: authenticity, integrity, timeliness, and orderliness of messages. That is, received data is handed over to the application layer only after passing the safety function module verification; data to be sent by the application layer is packaged by the safety communication module before being transmitted externally [11-12]. Therefore, we must ensure that every part adheres to strict safety regulations and take appropriate measures to protect them. This ensures a secure and reliable communication service environment. The position of the safety communication protocol in the safety communication model is shown in Figure 2.

DOI: 10.54254/2755-2721/87/20241523

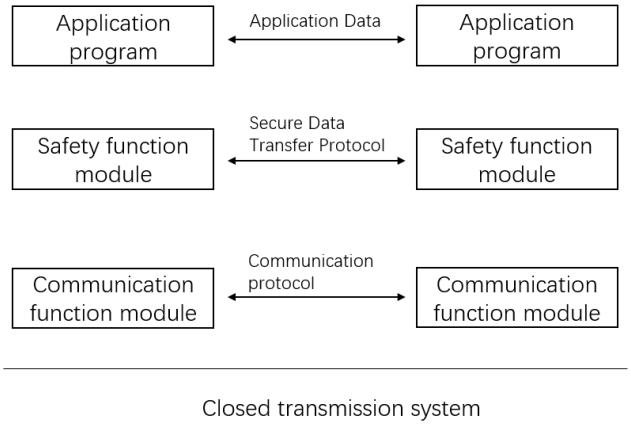

Figure 2. Safety Communication Model

To ensure the safety and reliability of the internal communication system, we divide it into three layers: the application layer, the safety protocol layer, and the communication base layer. Each layer defines the respective data formats. The application layer is responsible for processing the actual data required for interactions. The safety protocol layer ensures communication reliability and is designed according to the EN50159 standard. Finally, the communication base layer stores data in the inherent format specified by the device bus [13].

4. Communication Security Analysis of the Fully Electronic Computer Interlocking System

4.1. Analysis of Communication Security in the Fully Electronic Computer Interlocking System

Communication security in the fully electronic computer interlocking system ensures the safety of communication between different parts of the system to prevent unauthorized access, data leaks, or tampering [14]. To analyze the communication security of the fully electronic computer interlocking system, the following aspects are typically considered:

1. Encrypted Communication: Ensuring that data transmitted between different parts of the system is encrypted to prevent data theft or tampering. Common encryption algorithms include AES and RSA, which ensure data confidentiality during transmission.

2. Identity Authentication: Verifying the identities of users and devices within the system to ensure that the communicating parties are legitimate and trustworthy. Identity authentication mechanisms can prevent unauthorized access.

3. Access Control: Restricting user or device access to system data and functions to ensure that only authorized users can perform specific operations. Detailed access control effectively reduces potential security risks.

4. Firewalls and Intrusion Detection Systems: Setting up firewalls and intrusion detection systems within the system to monitor network traffic and behavior, quickly identifying potential attacks and preventing or alerting them in time.

5. Security Vulnerability Management: Regularly scanning and assessing the system for security vulnerabilities and promptly patching known vulnerabilities to maintain system security continuously.

6. Logging and Auditing: Recording operation logs within the system, including user logins, data access, and other behaviors, to trace and investigate security incidents when they occur.

7. Physical Security: Ensuring the physical security of the system's servers and network equipment to prevent unauthorized personnel from accessing and operating system hardware.

By comprehensively considering these factors, the communication security of the fully electronic computer interlocking system can be effectively guaranteed [15]. Additionally, continuously monitoring the latest developments and technologies in the security field and promptly adjusting and updating security strategies are crucial for maintaining system security.

DOI: 10.54254/2755-2721/87/20241523

4.2. Reliability Analysis of Communication in the Fully Electronic Computer Interlocking System

Reliability analysis of communication in the fully electronic computer interlocking system is crucial to ensure that data and commands are transmitted stably and efficiently during the communication process [16]. The following are common methods and strategies for evaluating and enhancing the reliability of communication in the fully electronic computer interlocking system:

1. Fault Analysis and Fault Tolerance Design: Analyzing and predicting potential communication failures in the system, designing fault tolerance mechanisms to handle communication failures, ensuring that the system can automatically switch to backup channels or recover to normal operation in case of issues.
2. Communication Link Quality Monitoring: Monitoring the quality and stability of each communication link in the system, including metrics such as delay, packet loss rate, and bandwidth utilization, to identify and adjust for communication problems promptly.
3. Data Integrity Check: Introducing verification mechanisms, such as CRC checks, during data transmission to ensure data integrity and prevent data corruption or tampering.
4. Redundant Communication Design: Implementing redundant communication paths or devices to achieve backup and redundancy in communication, enhancing system reliability and stability. If the primary communication path encounters problems, it can quickly switch to the backup path.
5. Network Topology Design: Designing a reasonable network topology to avoid single points of failure affecting the entire system's communication. Adopting distributed architectures and multi-path communication to improve the system's resistance to interference.
6. Communication Security Strategies: Implementing security measures such as encrypted communication, identity authentication, and access control to protect communication data security and prevent information leaks and attacks.
7. Regular Maintenance and Monitoring: Performing regular maintenance and monitoring of the system's communication equipment and network to detect and address potential issues promptly, ensuring the system's stability and reliability.

By comprehensively applying these measures and strategies, the communication reliability of the fully electronic computer interlocking system can be effectively enhanced, ensuring stable and efficient data transmission and command control during system operation [17-18].

5. Conclusion

In railway transportation, the train interlocking system ensures the safe and smooth passage of trains through intersections, shunting lines, and other sections to avoid accidents and collisions. The fully electronic computer interlocking system introduces modern electronic and computer technologies, improving the intelligence and safety of the railway transportation system while enhancing the efficiency and accuracy of train operations. Therefore, ensuring stable communication in the fully electronic computer interlocking system and providing a secure communication environment has become paramount in ensuring the normal operation of the railway control system.

References

[1] Ma Jun. Analysis and Research on Modern Railway Signal System [J]. SME Management and Technology (Late Issue), 2016, (02): 276.
[2] Duan Wu. Overview of the Development of Railway Station Interlocking in China [J]. Railway Communication Signal, 2019, 55(S1): 86-97.
[3] Fu Limin. Research on the Development and Application of Fully Electronic Interlocking [J]. Railway Communication Signal Engineering Technology, 2020, 17(03): 32-38.
[4] A. Nouri and J. Warmuth, "IEC 61508 and ISO 26262 - A Comparison Study," 2021 5th International Conference on System Reliability and Safety (ICSRS), Palermo, Italy, 2021: 138-142.
[5] DIN EN 50159-2011, Railway applications - Communication, signalling and processing systems - Safety-related communication in transmission systems; German version EN 50159:2010[s].

DOI: 10.54254/2755-2721/87/20241523

[6] Hassan Md Kamrul, Subramanian Kannan Bala, Saha Swapan, Sheikh M. Neaz. Behaviour of prefabricated steel-concrete composite slabs with a novel interlocking system – Numerical analysis [J]. Engineering Structures, 2021, 245.
[7] W. Fu, K. Wang, H. Feng and X. Ma, "Research on Computer Interlocking System with Interoperability Function," 2019 IEEE 2nd International Conference on Electronics and Communication Engineering (ICECE), Xi'an, China, 2019, 230-234.
[8] Lee J, Jung J. Verification and Conformance Test Generation of Communication Protocol for Railway Signaling Systems [J]. Computer Standards & Interfaces, 2007, 29(2): 143-151.
[9] Han Bingqian, Su Xiuyuan. Research on the Development and Application of Fully Electronic Interlocking System [J]. Railway Communication Signal Engineering Technology, 2022, 19(08): 92-96.
[10] Wang Yuetai, Wu Wen'ai. Analysis of Reliability and Safety of Computer Interlocking System [J]. Inner Mongolia Coal Economy, 2020, (05): 159.
[11] Zhang Hanbai. Communication Protocol Scheme and Security Impact in Fully Electronic Computer Interlocking System [J]. Digital World, 2019, (02): 26.
[12] H. Feng, J. Yu, X. Mo, M. Song and Y. Guan, "Research on All Electric Computer Interlocking System," 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 2021, 406-410.
[13] Xu Li, Su Siqi, Kuang Wenzhen. Design and Security Analysis of Communication Protocol for Fully Electronic Computer Interlocking System [J]. China Railway Science, 2012, 33(06): 83-87.
[14] Pinedo C, Aguado M, Lopez I. Modelling and Simulation of ERTMS for Current and Future Mobile Technologies [J]. International Journal of Vehicular Technology, 2015, 2015: 1-11.
[15] Franco D, Aguado M, Pinedo C. A Contribution to Safe Railway Operation: Evaluating the Effect of Electromagnetic Disturbances on Balise-to-BTM Communication in Railway Control Signaling Systems [J]. IEEE Vehicular Technology Magazine, 2021, 16(2): 104-112.
[16] Yin Qin, Zhang Liwei. Implementation Method of Secure Communication Protocol Based on Open Network [J]. Railway Communication Signal Engineering Technology, 2023, 20(01): 24-27+45.
[17] Feng Haonan. Design and Research of Fully Electronic Computer Interlocking System for Urban Rail Transit [J]. Journal of Railway Science and Engineering, 2021, 18(08): 2145-2155.
[18] Gao Yang, Luo Yangfan. Discussion on Security Requirements of Wireless Communication Protocol for High-speed Railway [J]. China Railway, 2022, (11): 123-128+134.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241515

Cognitive machine learning techniques for predictive maintenance in industrial systems: A data-driven analysis

Yinxuan Chai[1], Liangning Jin[2], Wentao Zhang[3,4,*]

[1]The University of Sydney, Sydney, Australia
[2]The University of Adelaide,South Australia, Australia
[3]The University of New South Wales, Sydney, Australia

[4]asderty67rtyasd@gmail.com
*corresponding author

Abstract. This paper delves into the intricate relationship between machine learning (ML) and data analysis, spotlighting the recent advancements, prevailing challenges, and emerging opportunities that underscore their integration. By conducting an extensive review of scholarly literature and real-world case studies, this article uncovers the synergistic potential of ML and data analysis, emphasizing their combined influence across diverse industries and domains. The exploration is framed around pivotal themes including algorithmic innovations, which are at the heart of ML's ability to transform vast and complex datasets into actionable insights. Moreover, the discussion extends to predictive modeling techniques, a cornerstone of data analysis that leverages historical data to forecast future trends, behaviors, and outcomes. Practical applications are scrutinized to demonstrate how the confluence of ML and data analysis is pioneering solutions in fields as varied as healthcare, where predictive analytics can save lives, to finance, where it is used to navigate market uncertainties. This paper also addresses the barriers to effective integration, such as data privacy concerns and the need for robust data governance frameworks. Through this comprehensive examination, the article sheds light on the rapidly evolving landscape of ML-driven data analysis, offering insights into how these technological advancements are reshaping research methodologies, industry practices, and societal norms.

Keywords: Machine Learning, Data Analysis, Integration, Advancements.

1. Introduction

In the contemporary landscape dominated by the deluge of data and rapid digitalization, the fields of machine learning (ML) and data analysis have ascended to the forefront, playing pivotal roles in navigating the complexities of modern data ecosystems. This introduction serves as a foundational primer, offering a comprehensive overview of ML and data analysis concepts to contextualize their significance in contemporary discourse. Machine learning, a subset of artificial intelligence (AI), encompasses a diverse array of algorithms and methodologies designed to empower computers to learn from data patterns and make predictions or decisions without explicit programming. [1] From supervised learning, where models are trained on labeled data, to unsupervised learning, where patterns and structures are inferred from unlabeled data, and reinforcement learning, where systems learn through trial and error, ML techniques underpin a broad spectrum of applications across industries and domains.

DOI: 10.54254/2755-2721/87/20241515

Concurrently, data analysis forms the bedrock of deriving actionable insights from data, involving the exploration, cleaning, and interpretation of datasets to uncover meaningful patterns, trends, and correlations. Through statistical methods, exploratory data analysis, and visualization techniques, data analysts illuminate the inherent structure and nuances within datasets, providing a foundation for informed decision-making and strategic planning. In essence, this article endeavors to delve into the intricate relationship between ML and data analysis, elucidating their complementary roles and synergistic potential in unlocking insights, driving innovation, and facilitating informed decision-making across myriad domains in the contemporary era of big data and digital transformation.

2. Algorithmic Innovations in Machine Learning

2.1. Deep Learning Architectures

Deep learning architectures, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have reshaped the landscape of pattern recognition and data modeling. CNNs excel in extracting hierarchical features from complex data, especially in image processing tasks, through convolutional layers and pooling operations. This hierarchical analysis enables CNNs to achieve state-of-the-art performance in tasks like object recognition, image classification, and segmentation, with applications spanning medical imaging, autonomous vehicles, and recommendation systems.

In sequential data analysis, RNNs play a pivotal role by capturing temporal dependencies among data points. Their recurrent connections allow them to maintain a memory of past inputs, making them ideal for tasks such as natural language processing (NLP), speech recognition, and time-series prediction. This capability to model sequential data enables RNNs to understand context and long-term dependencies, facilitating language translation, sentiment analysis, and speech synthesis across various applications like chatbots, virtual assistants, predictive text input, and music generation. [2]

More recently, transformer models have emerged as a significant advancement in deep learning architectures. Characterized by their attention mechanism and self-attention mechanisms, transformers selectively focus on relevant parts of the input sequence, allowing parallel processing and efficient learning of long-range dependencies.

2.2. Probabilistic Graphical Models

Probabilistic graphical models, such as Bayesian networks, offer a principled framework for representing and reasoning about uncertain relationships in data. Bayesian networks use directed acyclic graphs to model probabilistic dependencies between variables, enabling causal reasoning, probabilistic inference, and decision-making under uncertainty. In domains such as healthcare, Bayesian networks aid in disease diagnosis, treatment planning, and prognosis prediction by capturing complex relationships between symptoms, risk factors, and medical interventions. Additionally, Bayesian networks find applications in finance, where they facilitate risk assessment, portfolio optimization, and fraud detection by modeling dependencies between market variables, economic indicators, and financial instruments.

Hidden Markov models (HMMs) represent a class of probabilistic graphical models widely used for sequential data modeling and prediction. HMMs consist of a hidden state sequence and an observable sequence, where the hidden states represent latent variables capturing underlying dynamics, and the observable states represent observed data. [3] In speech recognition, HMMs model relationships between phonemes or words, enabling accurate transcription and speech synthesis. In genomic sequence analysis, HMMs model relationships between DNA or protein sequences, facilitating tasks such as sequence alignment, motif discovery, and gene prediction. HMMs have also found applications in natural language processing, where they model syntax, semantics, and discourse structure, enabling tasks such as part-of-speech tagging, named entity recognition, and parsing. As shown in Figure 1.

DOI: 10.54254/2755-2721/87/20241515

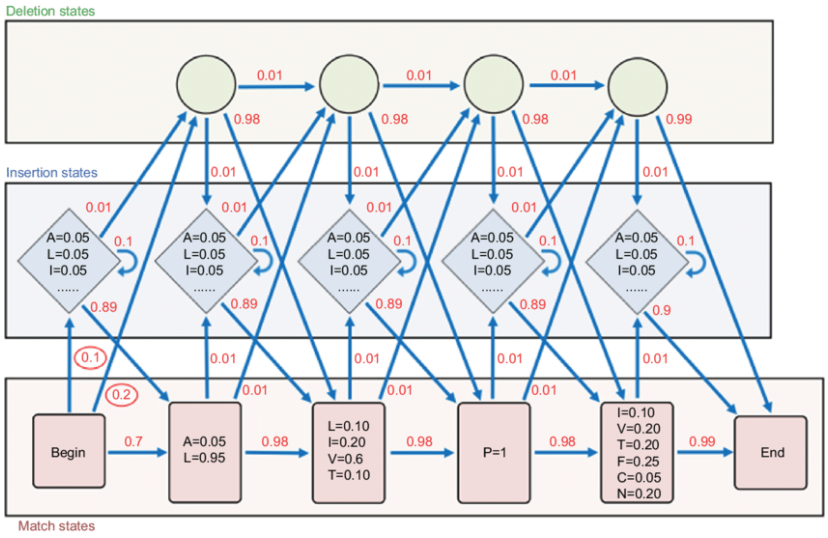

Figure 1. Predictive Modeling Techniques in Data Analysis (Source: ResearchGate)

2.3. Regression Analysis

Linear regression, a foundational technique in predictive modeling, aims to establish a linear relationship between independent and dependent variables. This method involves fitting a linear equation to observed data points, minimizing the sum of squared differences between predicted and actual values. Widely applied in fields such as finance and marketing, linear regression finds extensive use in forecasting stock prices, asset returns, and predicting consumer behavior and market trends. Its coefficients offer insights into the strength and direction of variable relationships, facilitating inference and decision-making. [4] Logistic regression extends these principles to model categorical outcomes, particularly binary events, by estimating the probability of occurrence based on predictor variables. Utilizing the logistic function, it maps the linear combination of predictors to a probability value between 0 and 1. Logistic regression is widely utilized in healthcare for disease diagnosis and risk prediction, as well as in marketing for customer segmentation and personalized campaigns. Ridge regression, as a regularization technique, addresses multicollinearity and overfitting issues in linear regression models by imposing a penalty on coefficient magnitudes. By augmenting the ordinary least squares objective function, ridge regression promotes smaller coefficient values, reducing model complexity. It is applied in finance for asset pricing models and risk management, and in healthcare for identifying disease biomarkers and treatment response prediction.

2.4. Time Series Forecasting

Autoregressive integrated moving average (ARIMA) models, exponential smoothing methods, and recurrent neural networks (RNNs) represent powerful techniques for time series forecasting across various domains. ARIMA models, renowned for their effectiveness in capturing trend and seasonality in data, decompose time series into autoregression (AR), differencing (I), and moving average (MA) components, allowing for the identification of temporal patterns and trends. Widely applied in finance for stock price prediction and portfolio optimization, as well as in energy forecasting for predicting electricity demand and supply fluctuations, ARIMA models offer valuable insights into future trends.

Exponential smoothing methods, including simple exponential smoothing (SES) and Holt-Winters exponential smoothing, provide effective solutions for forecasting time series data exhibiting exponential decay in trends and seasonality. These techniques assign exponentially decreasing weights to past observations, prioritizing recent data points. Commonly utilized in supply chain management for inventory forecasting and production planning, and in marketing for predicting sales and demand, exponential smoothing techniques play a pivotal role in optimizing resource allocation and strategic decision-making. [5]

DOI: 10.54254/2755-2721/87/20241515

3. Practical Applications of ML-Driven Data Analysis

3.1. Healthcare Informatics

Machine learning (ML) algorithms are revolutionizing healthcare informatics, particularly in the realm of personalized medicine. Leveraging patient data from electronic health records (EHRs), genetic information, and medical imaging data, these algorithms tailor treatment plans based on individual characteristics and medical history. By analyzing large and diverse datasets, ML-driven predictive models identify patient-specific risk factors, predict treatment responses, and recommend personalized interventions, ultimately improving patient outcomes and reducing healthcare costs. Additionally, ML-driven data analysis enhances disease diagnosis and biomarker identification through the integration of diverse data sources and advanced analytics techniques. By analyzing patterns in patient data, including clinical symptoms, genetic markers, and imaging findings, ML algorithms assist in early disease detection and classification. Furthermore, ML techniques enable the identification of disease biomarkers, facilitating the development of diagnostic tests and targeted therapies for various medical conditions, ranging from cancer to neurological disorders. [6] Moreover, ML-driven approaches optimize treatment strategies and provide clinical decision support by leveraging patient-specific data and evidence-based guidelines. These techniques analyze patient demographics, medical history, and treatment outcomes to identify optimal interventions and adjust treatment plans in real-time. Additionally, ML algorithms aid healthcare providers in prioritizing care delivery, predicting patient readmissions, and minimizing adverse events, ultimately enhancing the quality of care and patient safety in clinical settings.

3.2. Financial Analytics

Financial analytics harnesses machine learning (ML) algorithms to bolster various aspects of financial operations, including risk assessment, fraud detection, and algorithmic trading. In risk assessment, ML algorithms scrutinize extensive financial datasets, comprising historical market data, economic indicators, and portfolio performance metrics, to quantify and forecast diverse risks such as market risk, credit risk, and operational risk. Employing advanced predictive modeling techniques like time series analysis and Monte Carlo simulations, ML-driven risk assessment tools offer financial institutions valuable insights into potential risks and vulnerabilities, empowering proactive risk mitigation strategies and well-informed decision-making processes.

Furthermore, ML-driven data analysis plays a pivotal role in the detection and prevention of fraudulent activities within the financial sector. By scrutinizing transactional data, user behaviors, and network patterns, ML algorithms can pinpoint anomalous activities indicative of fraud or malicious intent. Leveraging techniques like anomaly detection, pattern recognition, and machine learning-based classification, financial institutions can effectively identify and thwart fraudulent transactions, unauthorized access attempts, and instances of identity theft. Such measures not only safeguard assets but also bolster trust in the integrity of the financial system.

3.3. Marketing Analytics

Machine learning (ML)-driven recommendation systems have revolutionized marketing analytics by providing highly personalized product recommendations based on consumer preferences and behaviors. These systems use collaborative filtering, content-based filtering, and hybrid approaches to analyze vast amounts of data. Collaborative filtering leverages user behavior to suggest products, with Amazon reporting that such recommendations drive 35% of their sales. Content-based filtering, used by Netflix, recommends items based on their attributes, saving the company approximately $1 billion annually by reducing churn. Hybrid systems, like those used by Spotify, combine both methods for enhanced accuracy. By analyzing historical purchase data, browsing history, and user interactions, these systems significantly boost customer engagement, increase cross-selling and upselling opportunities, and strengthen customer loyalty. McKinsey reports that personalized recommendations can increase sales by 10-30%. Walmart, for instance, processes over 2.5 petabytes of data every hour to refine its

DOI: 10.54254/2755-2721/87/20241515

algorithms. Overall, ML-driven recommendation systems are crucial for providing personalized experiences that drive engagement and sales in competitive markets.

4. Challenges and Opportunities in ML-Driven Data Analysis

4.1. Data Quality and Quantity

Data preprocessing and cleansing, along with feature engineering, selection, and data fusion/integration, collectively form the foundation of robust and effective machine learning (ML) analysis. These essential steps are pivotal in addressing the challenges posed by data quality and quantity in ML-driven data analysis.

Data preprocessing and cleansing involve a series of techniques aimed at ensuring that the data used for modeling are accurate, complete, and representative of the underlying phenomenon. Techniques such as outlier detection, missing value imputation, and normalization or scaling are employed to handle anomalies and inconsistencies in the data, thereby improving the quality of input data for ML models. Additionally, data augmentation methods, including synthetic data generation and oversampling, are utilized to alleviate data scarcity issues and enhance the quantity of data available for model training.

Data fusion and integration techniques enable the aggregation of heterogeneous data sources, thereby enhancing the quality and quantity of information available for ML-driven analysis. By merging data from multiple sources, including structured databases, unstructured text documents, and sensor streams, comprehensive datasets are created, capturing diverse aspects of the underlying problem. [7] This integration allows ML models to leverage complementary information, uncover hidden patterns, and improve predictive performance, ultimately facilitating more accurate decision-making and insights generation in complex and dynamic environments.

4.2. Interpretability and Explainability

In the realm of machine learning and artificial intelligence, the pursuit of interpretability and explainability is paramount for building trust and understanding the decisions made by these systems. Model-agnostic interpretability techniques, such as feature importance analysis and partial dependence plots, offer a broad perspective on the behavior of ML models, allowing stakeholders to dissect the influence of individual features on predictions. By delving into these insights, stakeholders can unravel hidden biases or limitations in the data or modeling process, empowering them to make informed decisions and mitigate risks effectively [8].

Complementing these techniques are Explainable AI (XAI) models, meticulously crafted to provide transparent and interpretable explanations for their predictions. Employing methodologies like rule-based models, decision trees, and symbolic reasoning systems, XAI models furnish human-readable explanations of ML predictions, elucidating the underlying logic and reasoning. This transparency not only fosters trust among stakeholders but also facilitates domain expert involvement and ensures regulatory compliance in critical sectors such as healthcare, finance, and criminal justice.

By integrating both model-agnostic interpretability techniques and XAI models into ML-driven analysis, stakeholders can create a robust framework for transparency, trust, and accountability in decision-making processes. [9] This holistic approach empowers stakeholders to delve deeper into the intricacies of ML models, address potential biases or limitations, and ultimately make more informed and ethical decisions across diverse domains. As shown in Table 1.

DOI: 10.54254/2755-2721/87/20241515

Table 1. Enhancing Transparency and Trust in Machine Learning through Interpretability

Techniques	Description
Model-agnostic Interpretability Techniques	Focus on understanding ML models independently of their architecture or learning algorithm. Examine contribution of individual features to model predictions. Valuable for identifying biases or limitations in data or modeling process.
Explainable AI (XAI) Models	Designed to provide transparent and interpretable explanations for predictions. Generate human-readable explanations of model decisions. Enhance trust, facilitate domain expert involvement, and support regulatory compliance.
Integration Approach	Integrates model-agnostic interpretability techniques and XAI models into ML-driven analysis. Enhances transparency, fosters trust, and ensures accountability in decision-making processes. Enables stakeholders to gain deeper insights into ML models and make more informed and ethical decisions.

5. Conclusion

In conclusion, machine learning and artificial intelligence stand as transformative technologies poised to revolutionize various industries and societal domains. Their applications span a wide spectrum, from healthcare and finance to marketing and beyond, offering unparalleled opportunities for innovation, efficiency, and progress. The impact of machine learning and AI is profound, with advancements leading to improved healthcare diagnostics and treatments, more accurate financial predictions and risk assessments, and highly targeted marketing strategies that enhance customer engagement.

However, amid the promising opportunities presented by these technologies, significant challenges and ethical considerations must be addressed. One of the foremost concerns is the potential for biases embedded in algorithms, leading to unfair outcomes and perpetuating existing social inequalities. Additionally, issues related to data privacy, security, and transparency require careful attention to ensure the responsible and ethical deployment of AI systems. Safeguarding sensitive information and ensuring transparency in AI decision-making processes are essential for building trust among users and stakeholders [10].

Furthermore, the ethical implications of AI-driven automation and job displacement need to be carefully managed to minimize adverse impacts on employment and socioeconomic stability. Efforts to upskill and reskill the workforce, coupled with policies that promote responsible AI adoption, can help mitigate these challenges and ensure a more equitable transition to a digitally driven future.

Contribution

Yinxuan Chai and Liangning Jin: Conceptualization, Methodology, Data curation, Writing- Original draft preparation, Visualization, Investigation.

References

[1] Chong, Edwin KP, Wu-Sheng Lu, and Stanislaw H. Zak. An Introduction to Optimization: With Applications to Machine Learning. John Wiley & Sons, 2023.

[2] Amini, Mahyar, and Ali Rahmani. "Agricultural databases evaluation with machine learning procedure." Australian Journal of Engineering and Applied Science 8.2023 (2023): 39-50.

[3] Copeland, Robert A. Enzymes: a practical introduction to structure, mechanism, and data analysis. John Wiley & Sons, 2023.

[4] Park, Minseok, and Nitya Prasad Singh. "Predicting supply chain risks through big data analytics: role of risk alert tool in mitigating business disruption." Benchmarking: An International Journal 30.5 (2023): 1457-1484.

DOI: 10.54254/2755-2721/87/20241515

[5] Ness, Stephanie, Nicki James Shepherd, and Teo Rong Xuan. "Synergy Between AI and Robotics: A Comprehensive Integration." Asian Journal of Research in Computer Science 16.4 (2023): 80-94.
[6] Rangineni, Sandeep, Divya Marupaka, and Arvind Kumar Bhardwaj. "An examination of machine learning in the process of data integration." International Journal of Computer Trends and Technology 71.6 (2023): 79-85.
[7] Satam, Heena, et al. "Next-generation sequencing technology: Current trends and advancements." Biology 12.7 (2023): 997.
[8] Iman, Mohammadreza, Hamid Reza Arabnia, and Khaled Rasheed. "A review of deep transfer learning and recent advancements." Technologies 11.2 (2023): 40.
[9] Aldoseri, Abdulaziz, Khalifa N. Al-Khalifa, and Abdel Magid Hamouda. "Re-thinking data strategy and integration for artificial intelligence: concepts, opportunities, and challenges." Applied Sciences 13.12 (2023): 7082.
[10] Ghazanfar, Shila, Carolina Guibentif, and John C. Marioni. "Stabilized mosaic single-cell data integration using unshared features." Nature biotechnology 42.2 (2024): 284-292.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241577

Applications and issues of artificial intelligence in the financial sector

Yanyu Chen

School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, Suzhou, China

Yanyu.Chen21@student.xjtlu.edu.cn

Abstract. This paper investigates the application of artificial intelligence in the financial sector, analyzing the existing technical, ethical, and legal issues, and proposing corresponding solutions. The research background highlights the widespread use of AI technology in the financial industry and the efficiency and cost benefits it brings. The research focuses on challenges related to data quality, feature engineering, model complexity, real-time capability, computational resources, and data privacy protection. The research method includes literature review and case analysis, revealing the applications of AI technology in stock prediction, risk management, trading strategy optimization, and customer service. The results indicate that effective data cleaning, automated feature engineering, model simplification and regularization techniques, the use of interpretability tools like LIME and SHAP, and the introduction of fairness evaluation standards can significantly enhance AI model performance and transparency. The conclusion points out that these measures can not only solve the current technical and ethical issues of AI in the financial sector but also promote the widespread application and standardization of AI technology in other fields.

Keywords: Artificial Intelligence, Financial Sector, Model Interpretability, Issues.

1. Introduction

In recent years, the rapid advancement of artificial intelligence technology has significantly transformed various sectors globally, leading to innovations that enhance efficiency, accuracy, and decision-making processes. As AI continues to evolve, its applications have expanded into multiple fields, including healthcare, transportation, manufacturing, and finance. Among these, the financial sector has particularly benefitted from AI's ability to automate difficult jobs, forecast market trends, and analyze enormous volumes of data. The importance of studying AI applications in the financial sector cannot be overstated. The financial sector, characterized by vast amounts of data and complex decision-making requirements, is particularly poised to benefit from AI technologies. Financial institutions have increasingly adopted AI-driven solutions to optimize operations, mitigate risks, and deliver personalized services to clients. The ability of AI to analyze large datasets, identify patterns, and make predictions offers significant advantages in areas such as fraud detection, algorithmic trading and document verification. These issues highlight the significance of conducting comprehensive research to understand the full impact of AI on the financial industry.

DOI: 10.54254/2755-2721/87/20241577

Despite the promising applications of AI in finance, there are substantial challenges that must be addressed to fully realize its potential. These include technical issues related to data quality and algorithm transparency, as well as ethical and legal concerns surrounding data privacy and financial liability. This research aims to explore the current applications of AI in finance, identify existing issues, and propose potential solutions. By reviewing the relevant literature, this study will provide a detailed overview of how AI technologies, such as machine learning, deep learning, and computer vision, are being utilized in financial services.

Furthermore, this study contributes significantly to the application of AI in the financial sector. The study identifies the key technical and ethical challenges facing the deployment of AI in the financial sector, providing insights into the complexities of integrating AI technologies into the industry. It also presents potential solutions and recommendations to address the identified challenges, aiming to guide financial institutions and policymakers to effectively utilise AI while mitigating the associated risks.

2. Artificial Intelligence Technology

2.1. Definition of Artificial Intelligence

Artificial intelligence can be summarized as a collection of analytical tools designed to mimic the cognitive functions of living beings. Over time, these tools have evolved into sophisticated instruments that address problems once considered difficult or impossible to solve [1]. AI involves the simulation of human intelligence in machines programmed to think and learn. These intelligent machines are capable of doing tasks like speech recognition, language translation, visual perception, and decision-making that normally need human intellect. AI systems are made to become more accurate at their duties, learn from mistakes, and adjust to new inputs.

2.2. Major AI Technologies

The advancement of AI technologies—encompassing machine learning, deep learning, and computer vision—has created numerous opportunities across various domains. Understanding these technologies and their applications is essential for maximizing AI's potential in the financial sector and beyond. This knowledge is crucial for effectively leveraging AI to enhance decision-making, improve efficiency, and drive innovation in diverse fields.

2.2.1. Machine Learning. Creating methods that let computers learn from data and make predictions is the main goal of the machine learning. It tackles the problem of developing automated systems that get better with use [2]. These models identify patterns and make decisions with minimal human intervention. As shown in Figure 1, key types include supervised learning, which uses labeled data; unsupervised learning, which finds patterns in unlabeled data; and reinforcement learning, which trains an agent through rewards and penalties.

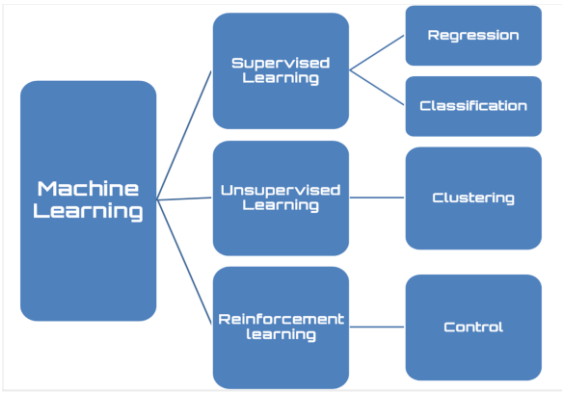

Figure 1. The key types of Machine Learning [3]

DOI: 10.54254/2755-2721/87/20241577

2.2.2. Deep Learning. Deep learning enables computational models with multiple processing layers to learn data representations with various levels of abstraction [4]. This approach has significantly advanced fields such as speech recognition, visual object recognition, object detection, drug discovery, and genomics. By employing multi-layered neural networks, deep learning effectively models complex relationships in data. Key components include Convolutional Neural Networks (CNNs) for image processing, a CNN architecture for handwritten digit recognition is shown in the figure 2. With such a multilayer architecture, CNN is able to extract and learn important features in an image, leading to efficient and accurate image classification.

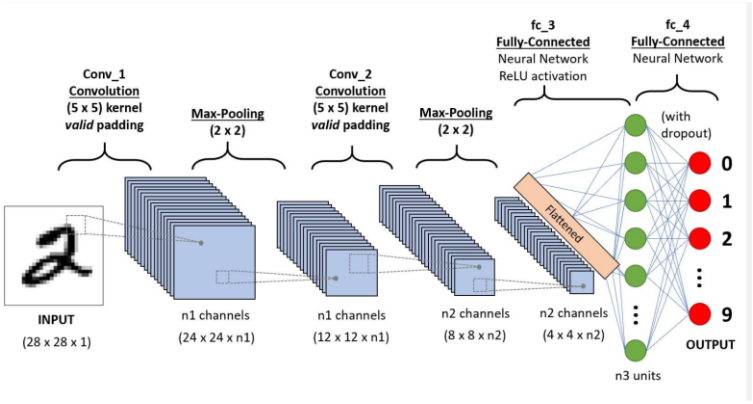

Figure 2. Flowchart of CNN [5]

2.2.3. Computer Vision. Computer vision enables computers to interpret visual data from the world. By establishing rules for pixel characteristics, relationships, and temporal changes, computer vision algorithms can automate the review of ecological images [6]. Applications include image classification, object detection and tracking, and image generation and enhancement. For example, the following figure 3 shows how a parallax image can be used for simple object detection, where objects in an image are distinguished into foreground and background by comparing pixel intensities with a set threshold. Such computer vision techniques can be used in autonomous driving to identify obstacles on the road, in medical imaging to help separate focal areas, and in surveillance to detect moving targets, thus enabling automated and intelligent image processing and analysis.

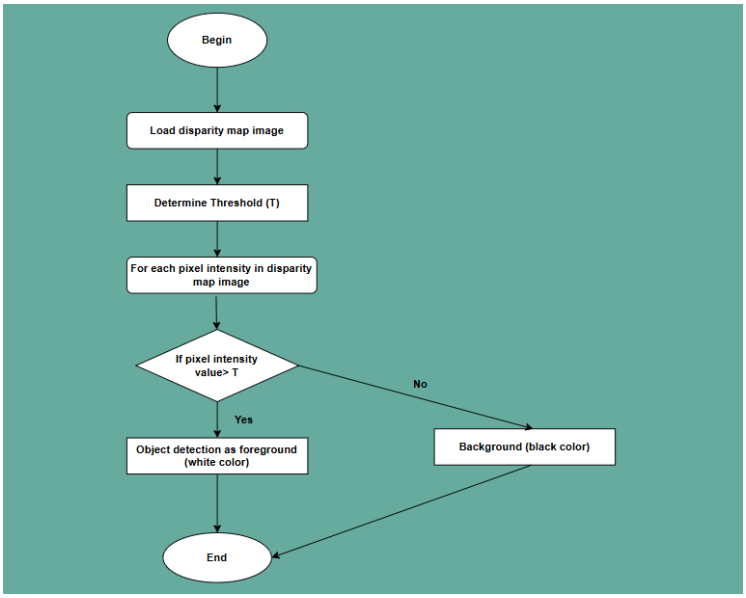

Figure 3. Flowchart for an Object Detection Process (Photo/Picture credit : Original)

DOI: 10.54254/2755-2721/87/20241577

3. Applications of AI in the Financial Sector

3.1. Machine Learning for Fraud Detection

Machine learning is revolutionizing fraud detection in the financial sector by utilizing sophisticated algorithms to analyze large volumes of transaction data and identify patterns indicative of fraudulent activities. Banks and financial institutions use supervised learning models to detect known fraud patterns, such as decision trees and random forest, which play crucial roles in identifying fraudulent transactions.

3.1.1. Decision Trees. A machine learning approach called a decision tree may be applied to applications involving regression and classification [7]. It simulates choices and their potential results, including as utility, resource costs, and outcomes from random events. Nodes and branches make up the tree structure, as seen in Figure 4. Based on eigenvalues, each internal node represents a decision point, and each branch, up to the output result, which is a leaf, indicates a decision outcome and links to other nodes.

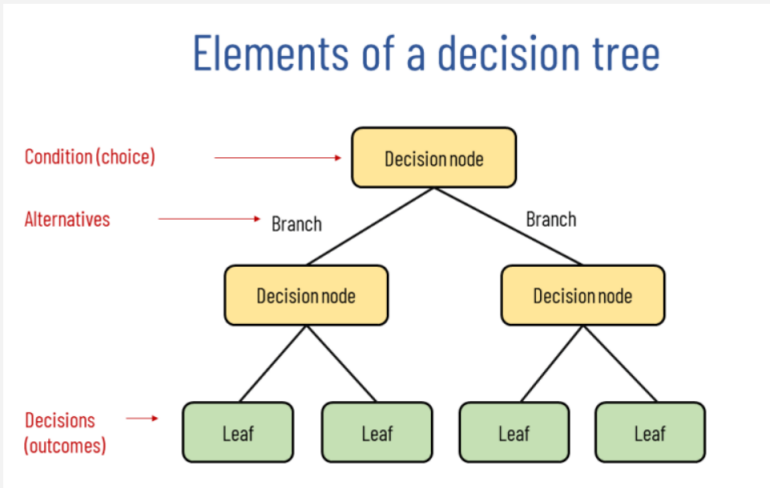

Figure 4. The structure of Decision Tree [8]

In the research of decision trees, major breakthroughs have been made in handling large amounts of features and data, enabling efficient classification and prediction. For example, decision trees can analyze various transaction attributes such as transaction amount, location, time, and user behavior. By training models on these attributes, decision trees can identify which combinations of features are more likely to indicate fraudulent activity. In practical applications, many financial institutions and payment platforms have begun using decision tree technology for fraud detection. For instance, e-commerce platforms like Amazon employ decision tree technology to protect both buyers and sellers. By analyzing users' browsing and purchase history, payment methods, and return behaviors, decision tree models can effectively identify potential fraudulent orders and take appropriate actions, such as temporarily freezing the transaction or conducting a manual review.

Similarly, credit card companies like Visa and MasterCard utilize decision tree technology to prevent fraudulent activities. These companies analyze historical transaction data to train decision tree models that can identify unusual transactions. Like Figure 5, if the system detects that a user's transaction location suddenly changes from one country to another and the transaction amount is substantial, the decision tree model may classify this transaction as high risk and promptly send a verification request to the user.

DOI: 10.54254/2755-2721/87/20241577

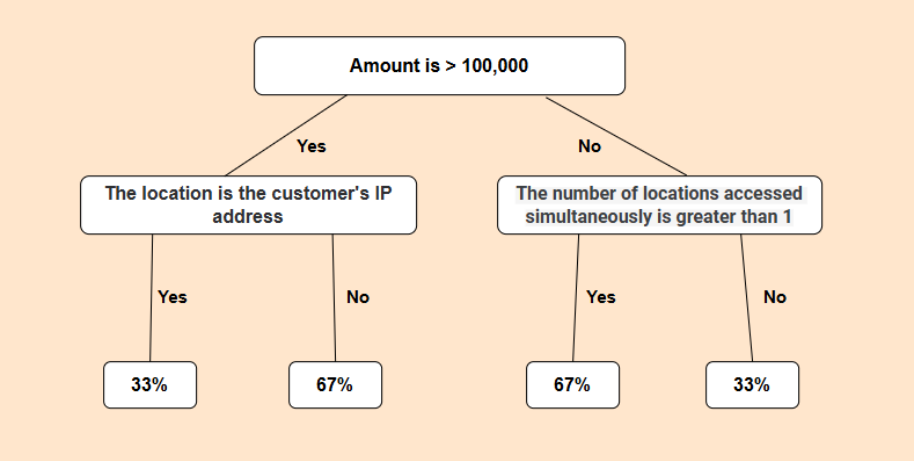

Figure 5. E-commerce Fraud Detection Decision Tree(Photo/Picture credit : Original)

3.1.2. Random Forest. During training, numerous decision trees are built using the random forest ensemble learning method, which then outputs the mean prediction (regression) or the mode of the classes (classification) for each individual tree [9]. It seeks to lower the chance of overfitting while enhancing forecast stability and accuracy. In order to ensure variety among the trees and strong overall performance, each tree in the forest takes into account a random selection of attributes and data points. Significant advancements have been made in the application of random forests, particularly in handling high-dimensional data and improving prediction accuracy. Random forests excel at managing large datasets with many features, as each tree in the forest is trained on a random subset of the data, which helps in capturing a wide range of patterns and relationships. This ensemble method is particularly effective in reducing variance and increasing model reliability.

In practical applications, JPMorgan Chase utilizes random forest algorithms to analyze its vast transaction data to identify potential fraudulent activities. The bank's system collects various features for each transaction, including transaction amount, frequency, location, time, and user's historical behavior patterns. The random forest model is trained on these features to identify abnormal patterns. For instance, if a user's transaction behavior suddenly changes from low-frequency, small-amount transactions to high-frequency, large-amount transactions, especially occurring in locations where the user does not usually appear, the model may flag these transactions as high risk. The bank then conducts further manual reviews on these high-risk transactions or directly contacts the user for confirmation. This approach enables JPMorgan Chase to promptly detect and prevent potential fraud, thereby protecting customers' assets.

3.2. Deep Learning for Algorithmic Trading

Algorithmic trading refers to the method of using computer algorithms to automatically buy and sell securities or other financial assets based on predetermined strategies and market conditions. It relies on high-speed computing and complex mathematical models to analyze market data, identify trading opportunities, and execute trades at the optimal times. The flowchart in Figure 6 clearly shows the basic working principle of algorithmic trading. Algorithmic trading uses software created by programmers to automate trading strategies. The software analyzes market conditions in real-time and executes trades quickly, reducing the workload for traders and allowing them to focus on strategy. This automation enhances efficiency and helps capture fast-moving market opportunities, making it a crucial tool in modern financial markets. Algorithmic trading can be applied to various financial markets, including stocks, futures, forex, and cryptocurrencies.

DOI: 10.54254/2755-2721/87/20241577

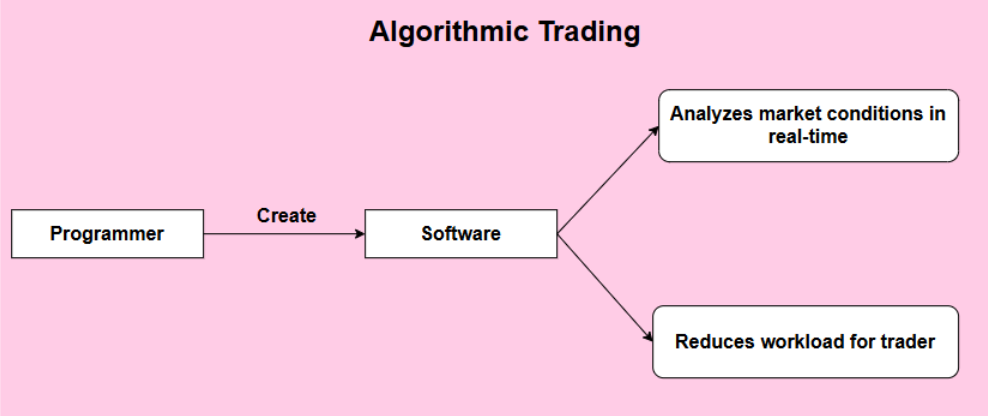

Figure 6. The process of Algorithmic Trading (Photo/Picture credit : Original)

In the context of algorithmic trading, deep learning techniques can analyze vast amounts of historical and real-time market data to predict future price movements and make trading decisions. Deep learning models, such as CNNs and Recurrent Neural Networks (RNNs), have proven particularly effective in this domain. In the following text, I will focus on CNN models to illustrate how deep learning specifically impacts algorithmic trading.

3.2.1. Convolutional Neural Networks (CNNs). CNNs, initially designed for image and spatial data analysis [10], have been successfully adapted for time-series data and financial chart analysis. CNNs capture spatial hierarchies in data through their unique convolutional layers, which contain filters—small matrices that move across the input data. Every filter is made to identify particular characteristics. A mathematical process known as convolution is carried out by the filters as they move across the data, creating a feature map that indicates the locations of features that have been identified. These layers apply filters to detect and learn complex patterns, such as trends and cycles in stock prices.

For instance, in trading applications, CNNs can be used to analyze candlestick charts or other visual representations of stock prices. Candlestick charts are widely used in stock trading to display the open, high, low, and close prices over a specific period. CNNs can identify shapes and patterns in stock price charts that often foreshadow significant price movements. For example, a trading algorithm can be trained to identify a "head and shoulders" pattern, which often indicates a reversal of an uptrend. When the algorithm identifies this pattern in real-time trading data, it can trigger a sell order to take advantage of an expected price drop. Similarly, the detection of a "double bottom" pattern can foreshadow an upcoming price increase, prompting the algorithm to place a buy order.

A practical application of CNNs in trading can be seen in quantitative hedge funds such as Renaissance Technologies and Two Sigma, which use machine learning models, including CNNs, to analyze large amounts of market data and identify trading opportunities. These funds use deep learning models like ResNet and Inception to detect complex patterns in price movements, allowing them to execute trades with greater accuracy and speed than traditional methods. Additionally, individual traders can use platforms such as MetaTrader or TradingView, which incorporate CNN-based indicators to help identify trading patterns. For example, traders using TradingView can use a ResNet-based plugin that automatically highlights potential "flag" patterns, which often indicate a continuation of a current trend, enabling traders to make informed decisions about entering or exiting a position.

By leveraging CNNs' ability to detect and learn complex patterns in stock prices, algorithmic trading systems can identify trading opportunities with higher precision and execute trades more effectively. This leads to more accurate and profitable trading strategies, enhancing efficiency and profitability in financial markets.

DOI: 10.54254/2755-2721/87/20241577

3.3. Computer Vision for Document Verification

With the advancement of artificial intelligence, computer vision applications in document verification have emerged as a significant research focus. Each year, numerous document verification tasks experience inefficiencies and frequent errors due to the constraints of manual processing. These constraints include human fatigue, negligence, inconsistency in subjective judgment, and slow processing speeds, leading to prolonged verification processes and a high likelihood of errors [11]. The introduction of computer vision technology can markedly improve the speed and accuracy of document verification. Computer vision employs automated processing and precise algorithms to rapidly analyze and verify large volumes of documents, thereby reducing human errors and enhancing overall efficiency.

The application of computer vision technology in document verification can be divided into two main parts: image processing and image analysis. Advanced image processing algorithms are used for pre-processing document images, including denoising and enhancement, which optimize image clarity and contrast for further analysis. In image analysis, optical character recognition (OCR) technology efficiently extracts text information from document images (Figure 7). Additionally, feature extraction and matching algorithms identify and verify key features in documents, such as signatures, seals, and security watermarks. These technologies significantly improve the accuracy and speed of document verification. The image below effectively demonstrates how an AI model, such as LeNet-5, processes and recognizes handwritten digits, showcasing the practical application of computer vision in understanding and verifying document content.

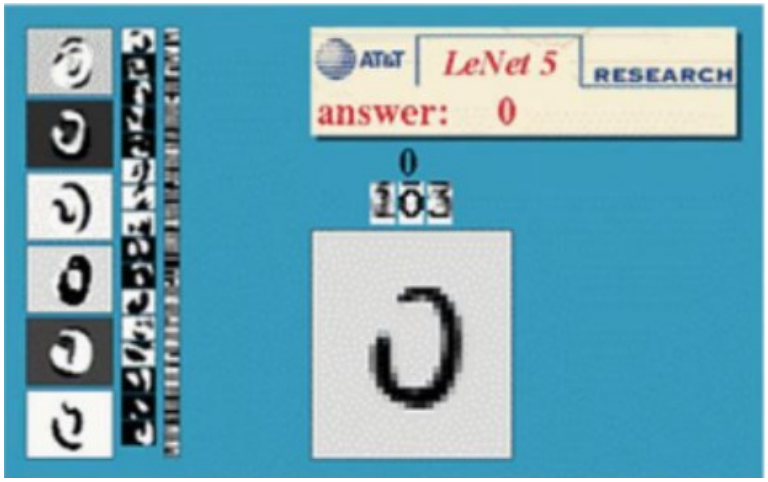

Figure 7. Optical Character Recognition (OCR) [12]

In practical applications, banks and financial institutions employ computer vision technology to automate the processing of vast amounts of customer documents, thereby reducing the workload and error rate associated with manual reviews. For example, Citibank and Bank of America utilize computer vision systems to manage account opening applications, loan documents, and customer identification documents. These systems swiftly scan and analyze documents, extract key information, and verify authenticity, enhancing processing efficiency and minimizing errors.

Moreover, government agencies leverage this technology to verify identity documents, ensuring the security and accuracy of the identity verification process. The U.S. Department of Homeland Security (DHS) uses computer vision technology to detect watermarks and security features in passports, ensuring the authenticity of passports and visas. Similarly, India's Aadhaar project employs computer vision technology to verify identity documents by scanning and matching biometric data on personal ID cards, guaranteeing the uniqueness and accuracy of each identity.

The advantages in government applications include enhanced security and reduced fraud risk. By automating the verification process, computer vision systems can detect sophisticated forgery attempts that might be missed by human inspectors. This technology also improves the efficiency of the

DOI: 10.54254/2755-2721/87/20241577

verification process, allowing agencies to handle a higher volume of documents without compromising accuracy.

On the downside, these systems can be vulnerable to adversarial attacks where malicious actors attempt to deceive the algorithm with specially crafted images. Additionally, the reliance on biometric data raises privacy concerns, as the collection and storage of such data must be handled with strict security measures to prevent unauthorized access and misuse.

Overall, computer vision technology not only enhances the accuracy and efficiency of document verification but also reduces reliance on manual labor, thereby lowering operational costs. Furthermore, it improves system security by reducing the risk of errors caused by human factors.

4. Issues in the Application of AI

4.1. Technical Issues

In the financial sector, AI technology is widely utilized for stock prediction, risk management, trading strategy optimization, and customer service. Despite the significant efficiency improvements and cost reductions brought by AI, its implementation still faces various technical challenges, limiting its full potential. Currently, data quality and acquisition issues are particularly prominent. For instance, in the stock market, trading data may contain noise and missing values, directly impacting the accuracy of model predictions. Moreover, due to privacy protection and regulatory restrictions, obtaining personal credit data is challenging, further hindering the training and application of credit scoring models and diminishing the value of AI in finance.

Data processing and feature engineering also present significant challenges. Financial data is usually high-dimensional, including numerous features such as market prices, trading volumes, and economic indicators. For example, in quantitative trading, analysts need to extract useful features from vast amounts of historical trading data to construct trading strategies. Processing this high-dimensional data and conducting effective feature selection and dimensionality reduction are crucial for enhancing model performance. However, finding relevant and useful features is often very complex, requiring the integration of domain expertise in finance, which complicates automation. The complexity of feature engineering not only affects the efficiency and effectiveness of model development but also imposes higher demands on the model's adaptability to different market environments.

The complexity and interpretability of models are also significant issues. The intricate and dynamic nature of financial markets necessitates models with high complexity and flexibility, yet overly complex models may lead to overfitting, reducing their ability to generalize to new data. For example, deep learning models used in hedge funds, while performing well on historical data, may not perform well under new market conditions, thus increasing investment risk. Furthermore, the decision-making processes of complex models are often hard to interpret, which not only affects user trust but also fails to meet regulatory requirements. Although complex models like deep learning offer performance advantages, their lack of transparency creates numerous obstacles in practical applications.

Finally, real-time capability and computational resources are critical issues for AI in the financial sector. Financial markets change rapidly, requiring models to have real-time analysis and decision-making capabilities. For instance, high-frequency trading systems demand data analysis and trading decisions within milliseconds, placing extremely high demands on computational resources. Additionally, training and running complex AI models require substantial computational and storage resources, posing a significant challenge for small to medium-sized financial institutions. The demand for high computational resources not only increases costs but also limits the broader application of AI technology. Data security and privacy protection are equally important; the sensitivity of financial data necessitates ensuring data security during storage, transmission, and processing to prevent leaks and tampering. For example, banks using AI for anti-money laundering detection must ensure the absolute security of customer data to prevent breaches and legal risks. Addressing these issues requires multi-faceted technical support and innovation to fully unleash the potential of AI in the financial sector.

DOI: 10.54254/2755-2721/87/20241577

4.2. Ethical and Legal Issues

In the financial sector, AI technology offers transformative potential but also raises significant ethical and legal issues. Despite the efficiency and innovation AI brings, its implementation must be carefully managed to address these concerns. Currently, issues related to bias and fairness are particularly prominent. For instance, AI models used in credit scoring may inadvertently perpetuate existing biases present in historical data, leading to discriminatory lending practices [13]. If the historical data shows a trend where certain demographics, such as minority groups, have been systematically denied loans or given higher interest rates, the AI model trained on this data might learn to replicate these patterns. As a result, even if two applicants have similar financial profiles, the AI system might unfairly favor applicants from non-minority backgrounds while disadvantaging those from minority groups, perpetuating a cycle of inequality and making it harder for affected individuals to access fair credit. Furthermore, the opacity of AI decision-making processes complicates the identification and correction of such biases, undermining fairness and potentially violating anti-discrimination laws.

There are additional difficulties with accountability and transparency. AI systems in finance frequently function as "black boxes," with opaque decision-making procedures. Insufficient transparency may result in a deficiency of responsibility, as it becomes challenging to assign blame for incorrect or detrimental judgments made by AI [14]. For example, it might be difficult to determine who is responsible and what options are available to impacted parties when an AI system wrongly rejects a loan application. In order to maintain responsibility and confidence, regulatory agencies are putting more and more pressure on AI systems to be open and to justify their decision-making procedures.

Lastly, the regulatory landscape for AI in finance is still evolving, creating legal uncertainties. Financial institutions must navigate a complex and rapidly changing regulatory environment that varies by region and jurisdiction. For example, the introduction of new AI-specific regulations or amendments to existing financial regulations can impact how AI systems are developed and deployed. Ensuring compliance with these evolving regulations requires significant resources and ongoing monitoring. Moreover, the lack of standardized regulations can lead to inconsistencies in AI governance, further complicating legal compliance and risk management.

5. Countermeasures and Recommendations

5.1. Technological Improvements and Innovations

To address data quality and acquisition issues, implementing effective data cleaning and preprocessing strategies, such as noise removal and missing value imputation, will ensure high-quality input data, thereby improving the accuracy of model predictions. Additionally, using data augmentation and synthesis techniques, such as Generative Adversarial Networks (GANs), can generate more high-quality training data, especially when personal credit data is limited, thus enhancing model performance. By implementing these strategies, noise and missing values in trading data will be effectively addressed, significantly improving data integrity and reliability, and consequently, model prediction accuracy. For the privacy and regulatory constraints on personal credit data, data augmentation techniques will provide an effective solution, ensuring that the model still receives sufficient training data under privacy protection conditions.

In data processing and feature engineering, the use of automated feature engineering can simplify the feature selection and engineering process, improve efficiency, and save time and resources. Encouraging close collaboration between financial experts and data scientists to extract and construct meaningful features will significantly enhance the model's applicability and stability. By implementing these measures, handling high-dimensional financial data will become more efficient, reducing the complexity of feature engineering, and significantly improving the efficiency and adaptability of model development. Particularly in quantitative trading, effective feature engineering will greatly improve the speed and quality of constructing trading strategies, enhancing the model's adaptability to different market environments.

DOI: 10.54254/2755-2721/87/20241577

To address model complexity and interpretability issues, using model simplification and regularization techniques can prevent overfitting and improve the model's generalization ability on new data. For example, regularization techniques can reduce model complexity, thereby enhancing its stability and reliability. Adopting interpretable models such as decision trees and linear regression, or using model interpretation tools like LIME and SHAP, can increase model transparency and meet regulatory requirements. By implementing these strategies, overfitting issues will be mitigated, the model's generalization ability will improve, and the transparency and interpretability of the model will be enhanced, addressing user trust and compliance issues. In applications like hedge funds, this will effectively reduce investment risk and improve model performance under new market conditions.

To tackle real-time capability and computational resources issues, developing and adopting efficient algorithms and optimization techniques can improve model computational efficiency, meeting the needs for real-time analysis and decision-making. For example, in high-frequency trading systems, fast algorithms can significantly enhance the speed of trading decisions. Utilizing cloud computing and edge computing technologies can provide powerful computing and storage resources, reducing the cost burden for small to medium-sized financial institutions while ensuring data security. By implementing these measures, the challenges posed by rapid changes in financial markets will be effectively addressed, the real-time requirements of high-frequency trading systems will be met, the pressure on computational resources will be greatly reduced, and data security will be ensured. This will make the training and operation of complex AI models more feasible and support the broader application of AI technology in more financial institutions.

5.2. Formulation of Relevant Regulations and Ethical Standards

To address bias and fairness issues, conducting data audits and ensuring diversity in training data before model training is crucial to prevent the perpetuation of historical biases. Introducing fairness evaluation standards, such as Demographic Parity and Equal Opportunity, can help correct biases during the model training process, ensuring model fairness. Demographic Parity requires that the distribution of the model's predictions be similar across different groups, meaning that the probability of a specific prediction (such as loan approval) should be the same for different groups (e.g., gender, race). Equal Opportunity focuses on the fairness of the model on true positive cases, ensuring that, in cases where a positive prediction is truly warranted, the probability of receiving a positive prediction is the same across different groups. By implementing these evaluation standards, the fairness of the model can be quantified, biases can be identified and corrected, ensuring that AI systems in financial applications do not lead to discriminatory practices, thus enhancing their ethical deployment.

To tackle transparency and accountability issues, using interpretability tools like LIME and SHAP can enhance model transparency, ensuring that decision-making processes are understandable and auditable. LIME (Local Interpretable Model-agnostic Explanations) is a model-agnostic explanation method that perturbs the input data locally, generates multiple similar data points, and observes how these perturbations affect the prediction results. This helps in constructing a simple, interpretable model that approximates the behavior of the complex model locally. LIME helps users understand the reasons behind specific predictions and reveals which features the model is sensitive to. SHAP (SHapley Additive exPlanations) is based on the Shapley value from cooperative game theory. It considers all possible combinations of features and their contributions to the prediction, calculating the marginal contribution of each feature. SHAP values provide a consistent measure of feature importance, explaining both global and local model behavior. By using these tools, the decision-making processes of AI models will become more transparent, increasing user trust and meeting regulatory requirements. This ensures that developers and users are accountable for model decisions and can provide effective remedies in case of errors or harm.

To address regulatory uncertainty issues, establishing a dedicated compliance team to continuously monitor and assess changes in relevant laws and regulations is vital to ensure that AI systems' development and deployment comply with the latest regulatory requirements. Additionally, participating in international and regional AI governance collaborations to promote the unification of

DOI: 10.54254/2755-2721/87/20241577

AI regulations and standards will reduce compliance difficulties caused by regulatory discrepancies. By implementing these measures, financial institutions can navigate the complex and evolving regulatory landscape more effectively, ensuring that their AI systems remain compliant and reducing the risks associated with legal uncertainties.

6. Conclusion

This research has examined the application of AI in the financial sector, focusing on its current uses, existing challenges, and potential solutions. The findings highlight the significant benefits of AI in enhancing efficiency, accuracy, and decision-making processes within financial institutions. Through effective data cleaning, automated feature engineering, model simplification, and the use of interpretability tools like LIME and SHAP, we can significantly improve the performance and transparency of AI models. These improvements address critical issues such as data quality, model complexity, and compliance with regulatory requirements.

The impact of this research extends beyond the immediate findings, offering valuable insights for improving AI applications across various sectors. For instance, the methods and solutions discussed can be adapted to enhance AI systems in healthcare, transportation, and manufacturing, where similar challenges of data quality and model interpretability exist. Additionally, the emphasis on ethical considerations and fairness evaluation standards provides a framework for addressing biases in AI models, ensuring more equitable and just applications of AI technology.

Looking to the future, this research underscores the importance of continuous innovation and collaboration in the development and deployment of AI. As AI technology evolves, it is crucial to stay ahead of emerging challenges through adaptive regulatory frameworks and robust ethical guidelines. Future studies should explore the integration of AI with other advanced technologies, such as blockchain and quantum computing, to further enhance its capabilities and applications. By fostering a multidisciplinary approach, we can ensure that AI technology continues to advance, driving innovation and positive change across all sectors of society.

References

[1] Gordon, B. M. 2011. Artificial Intelligence: Approaches, Tools, and Applications. Nova Science Publishers, Inc.
[2] Jordan, M. I., & Mitchell, T. M. 2015. Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.
[3] Ciaburro, G. 2017. MATLAB for machine learning. Packt Publishing Ltd.
[4] LeCun, Y., Bengio, Y., & Hinton, G. 2015. Deep learning. nature, 521(7553), 436-444.
[5] Saha, S. 2018. A comprehensive guide to convolutional neural networks—the ELI5 way. Towards data science, 15, 15.
[6] Weinstein, B. G. 2018. A computer vision for animal ecology. Journal of Animal Ecology, 87(3), 533-545.
[7] De Ville, B. 2013. Decision trees. Wiley Interdisciplinary Reviews: Computational Statistics, 5(6), 448-455.
[8] Image source: Yulia Kosarenko, 2021 How to Create Decision Trees for Business Rules Analysis," Why Change, November 13.
[9] Rigatti, S. J. 2017. Random forest. Journal of Insurance Medicine, 47(1), 31-39.
[10] O'shea, K., & Nash, R. 2015. An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.
[11] Sharma, N., Gupta, S., Mehta, P., Cheng, X., Shankar, A., Singh, P., & Nayak, S. R. 2022. Offline signature verification using deep neural network with application to computer vision. Journal of Electronic Imaging, 31(4), 041210-041210.
[12] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

DOI: 10.54254/2755-2721/87/20241577

[13] Griffith, M. A. 2023. AI Lending and the ECOA: Avoiding Accidental Discrimination. NC Bank. Inst., 27, 349.
[14] Uzougbo, N. S., Ikegwu, C. G., & Adewusi, A. O. 2024. Legal accountability and ethical considerations of AI in financial services. GSC Advanced Research and Reviews, 19(2), 130-142.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241527

Applications of stochastic processes and reinforcement learning in strategic decision support and personalized ad recommendation: An AIGC study

Qinxia Ma

Beijing University of Posts and Telecommunications, Beijing, China

maqinxia@126.com

Abstract. This paper investigates the integration of stochastic processes and reinforcement learning (RL) in strategic decision support systems (SDSS) and personalized advertisement recommendations. Stochastic processes offer a robust framework for modeling uncertainties and predicting future states across various domains, while RL facilitates dynamic optimization through continuous interaction with the environment. The combination of these technologies significantly enhances decision-making accuracy and efficiency, yielding substantial benefits in industries such as financial services, healthcare, logistics, retail, and manufacturing. By leveraging these advanced AI techniques, businesses can develop adaptive strategies that respond to real-time changes and optimize outcomes. This paper delves into the theoretical foundations of stochastic processes and RL, explores their practical implementations, and presents case studies that demonstrate their effectiveness. Furthermore, it addresses the computational complexity and ethical considerations related to these technologies, providing comprehensive insights into their potential and challenges. The findings highlight the transformative impact of integrating stochastic processes and RL in contemporary decision-making frameworks.

Keywords: Stochastic processes, reinforcement learning, strategic decision support systems, personalized advertisement recommendations

1. Introduction

The advent of artificial intelligence (AI) and machine learning (ML) has transformed numerous industries by enabling more accurate predictions, optimized decision-making, and personalized user experiences. Among various AI techniques, stochastic processes and reinforcement learning (RL) have emerged as powerful tools for strategic planning and decision support. Stochastic processes incorporate randomness into mathematical models, allowing for the prediction and analysis of systems that evolve over time. This capability is particularly valuable in environments characterized by uncertainty and variability, such as financial markets, supply chains, and healthcare. Reinforcement learning focuses on how agents can learn to make optimal decisions through interactions with their environment, maximizing cumulative rewards. By combining these two approaches, organizations can develop adaptive decision-making frameworks that are resilient to changes and capable of optimizing outcomes in real time. The integration of stochastic processes and RL offers a compelling solution to the complexities of modern strategic decision-making. In financial services, stochastic models can forecast

DOI: 10.54254/2755-2721/87/20241527

market trends while RL algorithms optimize trading strategies to maximize returns. In logistics, these technologies can predict traffic patterns and dynamically adjust routing and scheduling, reducing operational costs and improving service delivery. Healthcare applications include patient outcome prediction and treatment plan optimization, enhancing personalized care and improving patient outcomes [1]. Retailers leverage these AI techniques for demand forecasting and inventory management, leading to increased sales and optimized stock levels. In manufacturing, predictive maintenance systems use stochastic models and RL to anticipate equipment failures and schedule timely repairs, minimizing downtime and maintenance costs. Despite their potential, the adoption of stochastic processes and RL faces challenges such as computational complexity, data quality, and ethical considerations, which must be addressed to fully realize the benefits of these technologies.

2. Theoretical Foundations

2.1. Stochastic Processes in Decision Support

Stochastic processes are mathematical models that incorporate randomness and are used to predict and analyze systems that evolve over time. In the context of decision support, these processes enable the modeling of uncertainties and variabilities inherent in strategic planning. For instance, in supply chain management, stochastic models can forecast demand fluctuations, helping businesses to optimize inventory levels and minimize costs. By incorporating random variables and probabilistic distributions, organizations can develop more resilient strategies that account for potential risks and uncertainties, ultimately leading to better-informed decisions. These models are also essential in financial risk management, where they help in assessing the volatility of asset prices and the likelihood of various economic scenarios. The use of stochastic processes in these areas ensures that decisions are based on a comprehensive analysis of possible outcomes, reducing the impact of unforeseen events. Figure 1 illustrates the impact of stochastic processes in decision support for supply chain management and financial risk management [2].

Figure 1. Impact Of Stochastic Processes In Decision Support

2.2. Reinforcement Learning Mechanisms

Reinforcement learning (RL) is a dynamic area of AI that focuses on how agents ought to take actions in an environment to maximize cumulative rewards. In RL, agents learn to make decisions by interacting with their environment, receiving feedback in the form of rewards or penalties. This learning paradigm is particularly effective in scenarios where the optimal strategy is not immediately obvious and must be discovered through exploration. For example, in automated trading systems, RL algorithms can learn to adjust trading strategies based on market conditions, thereby maximizing profits. The ability of RL to adapt and optimize decision-making processes in real-time makes it a powerful tool for strategic

DOI: 10.54254/2755-2721/87/20241527

planning. Furthermore, RL has applications in robotics, where it enables machines to learn complex tasks through trial and error, improving their performance over time. The flexibility and adaptability of RL make it suitable for a wide range of applications, from game playing to industrial automation [3]. One fundamental equation in RL is the Bellman equation, which is central to many RL algorithms:

$$Q^{\pi}(s,a) = E_{\pi}[r_t + \gamma \cdot Q^{\pi}(s_{t+1}, a_{t+1}) | s_t = s, a_t = a] \quad (1)$$

Where:

$Q^{\pi}(s,a)$ is the action-value function, representing the expected return (sum of rewards) after taking action aa in state ss and following policy π. rt is the reward received after taking action aa in state ss. γ is the discount factor, which determines the importance of future rewards ($0 \leqslant \gamma \leqslant 1$). s_{t+1} and a_{t+1} are the state and action at the next time step. The Bellman equation can be used to iteratively improve the policy π by updating the action-value function Q. This iterative process is a key mechanism in RL algorithms like Q-learning and Deep Q-Networks (DQN).

2.3. Synergy Between Stochastic Processes and RL

The integration of stochastic processes and reinforcement learning creates a powerful synergy for decision support systems. Stochastic processes provide a structured way to model uncertainties and predict outcomes, while RL algorithms utilize these models to learn and optimize strategies over time. This combination allows for the development of adaptive decision-making frameworks that can respond to changing conditions and uncertainties. For instance, in personalized advertisement recommendation, stochastic models can predict user behavior patterns, and RL can adjust ad placements based on these predictions to enhance user engagement and conversion rates [4]. This synergy is also evident in healthcare, where predictive models can anticipate patient outcomes, and RL can optimize treatment plans based on these predictions. By combining these approaches, organizations can achieve a higher level of precision and efficiency in their decision-making processes, leading to improved outcomes and greater operational effectiveness.

3. Practical Implementations

3.1. Strategic Decision Support Systems

Table 1. Strategic Decision Support Systems (SDSS) applications

Industry	Application	Stochastic Model	Reinforcement Learning	Benefits
Financial Services	Market trend forecasting and trading strategy development	ARIMA, GARCH	Q-learning, Deep Q Networks	Maximized returns, reduced financial risk
Logistics	Routing and scheduling optimization based on traffic predictions	Poisson Process, Markov Chains	Policy Gradient Methods, Actor-Critic Models	Reduced operational costs, improved service delivery
Healthcare	Patient outcome prediction and treatment plan optimization	Bayesian Networks, Hidden Markov Models	Deep Q Networks, Actor-Critic Methods	Personalized treatment plans, improved patient outcomes
Retail	Demand forecasting and inventory management	Exponential Smoothing, Monte Carlo Simulation	Multi-Armed Bandit, Deep Q Networks	Optimized inventory levels, increased sales
Manufacturing	Predictive maintenance and repair scheduling	Weibull Distribution, Gamma Process	Deep Q Networks, Policy Gradient Methods	Reduced downtime, lower maintenance costs

DOI: 10.54254/2755-2721/87/20241527

Strategic decision support systems (SDSS) leverage AI technologies to assist organizations in making informed decisions. By integrating stochastic processes and RL, these systems can analyze vast amounts of data, identify trends, and optimize decision-making strategies. For example, in financial services, SDSS can use stochastic models to forecast market trends and RL to develop trading strategies that maximize returns. The combination of these technologies enables businesses to navigate complex environments with greater confidence, making more accurate and strategic decisions. In logistics, SDSS can optimize routing and scheduling by predicting traffic patterns and adjusting plans in real-time. This results in reduced operational costs and improved service delivery. The versatility of SDSS in various industries underscores its importance as a tool for enhancing organizational performance and competitiveness [5]. Table 1 includes various industries, their respective applications, the stochastic models and reinforcement learning methods used, and the benefits gained from implementing these systems.

3.2. Personalized Advertisement Recommendations

Personalized advertisement recommendation systems aim to deliver tailored ad content to individual users based on their preferences and behavior. By utilizing stochastic processes to model user interactions and RL to optimize ad placements, these systems can significantly enhance ad relevance and user engagement. For instance, e-commerce platforms can use stochastic models to predict user purchase patterns and RL algorithms to recommend products that align with these patterns. This personalized approach not only improves user experience but also increases the likelihood of conversion, making advertising efforts more effective [6]. Additionally, in streaming services, personalized recommendations can enhance viewer satisfaction by suggesting content that aligns with their viewing history and preferences. This leads to increased user retention and engagement, driving higher revenue for the service providers. The ability to deliver highly relevant content is a key advantage of personalized recommendation systems, making them an essential component of modern marketing strategies.

3.3. Case Studies in Industry

The practical application of stochastic processes and reinforcement learning (RL) has led to significant advancements across various industries. In the financial sector, a leading investment firm implemented an SDSS that combined ARIMA models for market trend forecasting with Deep Q Networks (DQN) for trading strategy optimization. This integration reduced financial risks and enhanced trading performance. In logistics, a global company adopted an SDSS using Poisson processes for traffic prediction and Policy Gradient Methods for dynamic routing and scheduling, resulting in lower operational costs and improved service delivery. In healthcare, a hospital network utilized Bayesian Networks and Actor-Critic Methods to predict patient outcomes and optimize treatment plans, leading to better patient care. A prominent retail chain used Exponential Smoothing models and Multi-Armed Bandit algorithms for demand forecasting and inventory management, increasing sales and reducing stockouts. In manufacturing, a predictive maintenance system employing Weibull Distribution models and Deep Q Networks minimized downtime and maintenance costs by accurately predicting equipment failures and scheduling timely repairs [7]. These case studies demonstrate the versatility and effectiveness of integrating stochastic processes and RL in various industries, driving innovation, efficiency, and improved decision-making.

4. Challenges and Limitations

4.1. Computational Complexity

One of the primary challenges associated with implementing stochastic processes and RL is the computational complexity involved. Both technologies require significant computational resources to process large datasets and perform complex calculations. This complexity can be a barrier to adoption, particularly for smaller organizations with limited resources. Additionally, the need for specialized expertise to develop and maintain these systems further complicates their implementation. Addressing

DOI: 10.54254/2755-2721/87/20241527

these challenges requires advancements in computational efficiency and the development of user-friendly tools that can simplify the adoption process. High-performance computing infrastructure and cloud-based solutions can mitigate some of these challenges, providing scalable resources for organizations to leverage these advanced AI techniques [8]. Continuous research and development in this area are essential to making these technologies more accessible and practical for widespread use. Table 2 provides a concise overview of the challenges and solutions related to computational complexity in the context of stochastic processes and RL, highlighting the benefits of addressing these challenges.

Table 2. Computational Complexity Challenges And Solutions

Challenge	Impact	Solution	Benefits
High Computational Requirements	Increased operational costs	High-performance computing infrastructure	Reduced costs, faster processing
Large Dataset Processing	Long processing times	Efficient data processing algorithms	Timely data analysis, enhanced performance
Complex Calculations	High error rates without precise calculations	Advanced mathematical techniques	Accurate results, improved decision making
Need for Specialized Expertise	Difficulty in system development and maintenance	User-friendly development tools	Simplified development, easier maintenance
Resource Limitations for Smaller Organizations	Barrier to adoption	Cloud-based scalable resources	Access to advanced AI techniques, increased adoption

4.2. Data Quality and Availability

The effectiveness of stochastic processes and RL heavily depends on the quality and availability of data. Inaccurate or incomplete data can lead to suboptimal models and flawed decision-making processes. Ensuring data integrity and access to relevant datasets is crucial for the successful implementation of these technologies. Furthermore, the dynamic nature of real-world environments means that data must be continuously updated to reflect current conditions. Organizations must establish robust data management practices and invest in data acquisition and maintenance to overcome these challenges [9]. Data governance frameworks and advanced data cleaning techniques can enhance data quality, ensuring that AI models are built on reliable and accurate information. Collaboration with data providers and the development of industry-specific data standards can also improve data availability, supporting the effective use of AI technologies.

4.3. Ethical and Privacy Concerns

The use of AI technologies in decision support and personalized advertising raises ethical and privacy concerns. The collection and analysis of user data for personalized recommendations can infringe on privacy rights, and biased algorithms can lead to unfair or discriminatory outcomes. Ensuring transparency, fairness, and accountability in AI systems is essential to addressing these concerns. Organizations must implement ethical guidelines and comply with regulatory standards to protect user privacy and promote the responsible use of AI technologies. Ethical AI practices involve designing algorithms that are explainable and auditable, allowing stakeholders to understand and trust the decision-making processes [10]. Regular audits and impact assessments can help identify and mitigate potential biases, ensuring that AI systems operate fairly and ethically. By prioritizing ethical considerations, organizations can build trust with users and create AI systems that benefit society as a whole.

DOI: 10.54254/2755-2721/87/20241527

5. Conclusion

The integration of stochastic processes and reinforcement learning in strategic decision support systems and personalized advertisement recommendations has the potential to revolutionize traditional business practices. By modeling uncertainties and optimizing decision-making processes in real time, these technologies enhance organizational performance, reduce costs, and improve service delivery. Case studies across various industries demonstrate the practical benefits and versatility of these AI techniques, showcasing their ability to drive innovation and efficiency. However, challenges such as computational complexity, data quality, and ethical considerations must be addressed to ensure successful implementation. Enhancing computational efficiency, improving data management practices, and developing robust ethical AI guidelines are essential steps towards making these technologies more accessible and practical for widespread use. Looking to the future, advancements in quantum computing, more sophisticated algorithms, and enhanced data processing capabilities promise to further amplify the impact of stochastic processes and RL. Additionally, interdisciplinary collaboration will be crucial in overcoming current limitations and exploring new applications. As research and development in this field continue to evolve, the potential for these technologies to achieve even greater precision and efficiency in decision-making frameworks is vast. Embracing these advancements will enable organizations to stay competitive and innovative in an increasingly complex and dynamic environment.

References

[1] Wang, Fuzhang, et al. “Artificial intelligence and stochastic optimization algorithms for the chaotic datasets.” FRACTALS (fractals) 31.06 (2023): 1-14.

[2] Albergo, Michael S., Nicholas M. Boffi, and Eric Vanden-Eijnden. “Stochastic interpolants: A unifying framework for flows and diffusions.” arXiv preprint arXiv:2303.08797 (2023).

[3] Faulwasser, Timm, et al. “Behavioral theory for stochastic systems? A data-driven journey from Willems to Wiener and back again.” Annual Reviews in Control (2023).

[4] Dutordoir, Vincent, et al. “Neural diffusion processes.” International Conference on Machine Learning. PMLR, 2023.

[5] Rosati, Riccardo, et al. “From knowledge-based to big data analytic model: a novel IoT and machine learning based decision support system for predictive maintenance in Industry 4.0.” Journal of Intelligent Manufacturing 34.1 (2023): 107-121.

[6] Higgins, Oliver, et al. “Artificial intelligence (AI) and machine learning (ML) based decision support systems in mental health: An integrative review.” International Journal of Mental Health Nursing 32.4 (2023): 966-978.

[7] Ali, Md Ramjan, Shah Md Ashiquzzaman Nipu, and Sharfuddin Ahmed Khan. “A decision support system for classifying supplier selection criteria using machine learning and random forest approach.” Decision Analytics Journal 7 (2023): 100238.

[8] Valluri, Durga Deepak. “Exploring cognitive reflection for decision-making in robots: Insights and implications.” International Journal of Science and Research Archive 11.2 (2024): 518-530.

[9] Rolf, Benjamin, et al. “A review on reinforcement learning algorithms and applications in supply chain management.” International Journal of Production Research 61.20 (2023): 7151-7179.

[10] Hasan, MD Rokibul. “Addressing Seasonality and Trend Detection in Predictive Sales Forecasting: A Machine Learning Perspective.” Journal of Business and Management Studies 6.2 (2024): 100-109.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241574

Research on recurrent neural network recommendation algorithm based on time series

Danping Qiu

Guangdong Baiyun College, China

15717075089@139.com

Abstract. In the era of the Internet, all people's behaviors on the Internet will be stored in the server in the form of data, which leads to the image level growth of data today. In the context of big data, how to effectively analyze various existing data to obtain the necessary information is an urgent challenge that various industries need to overcome. Recommendation algorithms are one of them, mainly using existing data to recommend information of interest to users. In traditional recommendation algorithms, collaborative filtering recommendation algorithms encounter difficulties such as cold start and data sparsity. In order to better explore data features, deep learning algorithms have begun to be applied. Recurrent neural networks can not only learn input data but also perform self-learning, which can better extract features between data and improve the accuracy of recommendations. However, the information that users are interested in is greatly influenced by time, in order to improve the accuracy of recommendations. This study investigates the recursive neural network recommendation algorithm with added time series, and experiments have shown that this recommendation algorithm can indeed improve the accuracy of recommendations more accurately.

Keywords: Recurrent neural network, Time series, Recommendation algorithm, Deep learning.

1. Introduction

With the in-depth development of information technology, the Internet has brought more and more convenience to our lives. At the same time, all operations on the Internet will be stored in the form of data, so the amount of data generated is also increasing exponentially. Behind such a vast amount of data, how to make good use of it and extract useful information through the data is also a topic that needs further research, such as the birth of recommendation algorithms to quickly recommend items of interest to users. The traditional recommendation algorithm is best known for collaborative filtering. In order to improve the accuracy of recommendations, scholars have made various improvements to the recommendation algorithm. Although it can effectively improve accuracy and solve some challenges such as insufficient data in the initial stage, there are still many challenges waiting to be overcome. In 2010, deep learning was promoted, and it belongs to a new type of computing model in the field of artificial intelligence[1]. It has become familiar to people and has been effectively used in practical scenarios and results, achieving good results in image and speech processing. Due to its excellent feature learning and data processing capabilities, deep learning can analyze the connections between data at a deeper level. Therefore, since the 2016 seminar on recommendation algorithms based on deep learning,

DOI: 10.54254/2755-2721/87/20241574

many experts and scholars have also conducted research on recommendation algorithms from the perspective of deep learning[2].

Lingyuan Dou proposed a collaborative filtering recommendation algorithm that integrates label features and temporal context[3]. This algorithm utilizes the correlation between information to reveal the relationship between individuals and objects and embeds it into neighbor based methods; In this way, the difficulties of information cold start in traditional recommendation systems can be successfully solved, thereby improving their accuracy level. However, how to incorporate time and situational factors into this algorithm is a tricky issue: as individual preferences and tendencies undergo variable changes over time. However, the introduction of temporal context is a complex issue, as user interests and preferences are dynamically changing over time. The paper mentions considering the temporal context factor of user ratings, but accurately capturing and utilizing this dynamic change remains a challenge. In addition, the impact of different time scales (such as days, weeks, months, etc.) on recommendation results may also vary, and further research is needed. Sun Guangfu proposed a collaborative filtering recommendation algorithm based on temporal behavior in his research, and developed a collaborative filtering recommendation method based on temporal behavior under this architecture[4]. This novel method aims to extract structured connections from the temporal consumption data of users and products, thereby establishing new associations between them. Compared to traditional collaborative filtering techniques, this method focuses more on analyzing the consumption time of users and products, in order to reveal their mutual influence. This method helps to solve the problem of interest drift and can better capture the dynamic preference changes of users. This algorithm has also been tested on the Douban recommendation dataset, and the results show that compared to traditional social network information and tag information recommendation methods, it can more accurately predict the true rating of users, thereby improving the accuracy of recommendations. However, this algorithm faces sparsity in consumer network graphs and cold start issues for users and products. These issues may affect the accuracy and efficiency of recommendation systems, especially when new customers or services are incorporated into this framework; Due to insufficient data interaction information as a basic support, conventional methods are difficult to provide useful solutions. To overcome this problem, Deng Cunbin proposed a novel strategy that integrates dynamic collaborative filtering and deep learning methods: by introducing time factors and utilizing deep learning to deal with problems that exist in ordinary patterns such as insufficient information, initial uncertainty, and the inability to consider differences in consumer demand that arise with time and environmental changes[5]. This algorithm utilizes dynamic collaborative filtering algorithm to integrate temporal features, and then uses higher-level machine intelligence technology to extract more relevant attributes in order to enhance its performance ability. Specifically, they used two algorithms, CNN and MLP, to obtain deep level representation features. Secondly, the dynamic collaborative filtering algorithm and deep learning model were combined to form a hybrid recommendation algorithm, and the rating prediction function and loss function of the dynamic collaborative filtering algorithm were improved. Finally, experiments on the MovieLens dataset have demonstrated that this method improves the accuracy of movie rating prediction. However, this algorithm integrates multiple models and technologies, including dynamic collaborative filtering, convolutional neural networks (CNN), and multi-layer perceptrons (MLP), and its structure may be relatively complex, resulting in reduced interpretability of the algorithm. In practical applications, this may affect the user's understanding and trust in the recommendation results. This study constructs a time series recurrent neural network recommendation algorithm based on previous literature discussions. This algorithm mainly identifies the user's points of interest in the time series and uses deep learning techniques to train the user's features, thereby improving the accuracy of recommendations.

2. Introduction to Recurrent Neural Networks

Recurrent neural network is a deep learning model that captures temporal dynamics by introducing a cyclic structure when processing sequential data[6]. This type of network can store previously inputted information and combine it with the current input for prediction or classification. RNN is different from traditional feedforward neural networks in that it has internal state or memory capabilities, which enable

DOI: 10.54254/2755-2721/87/20241574

it to process long time series and adapt to time changes. The core feature of recursive neural networks is the cyclic connections of their hidden layers. This architecture indicates that the formation of network output is not only dependent on the data input of the existing input layer, but also influenced by the data loop input of the previous hidden layer and system conditions. Therefore, this feature makes it very suitable for performing complex tasks that require consideration of time, such as voice recognition, text generation, or automatic decoding. During the training phase, RNN uses backpropagation algorithm (BPTT) to adjust and optimize parameters[7]. However, due to issues such as vanishing gradients or exploding data, standardized RNNs may pose challenges for long-term information transmission. To overcome this challenge, scholars have developed many improved versions of RNNs, such as LSTM and GRU, which introduce a technology called "gated loop unit" to effectively manage information flow and more accurately distinguish factors that affect the future. In summary, recurrent neural networks are a powerful tool for deep learning, especially effective for tasks that require consideration of time series. By introducing a loop structure and an improved gating mechanism, RNN can effectively capture temporal dynamics and solve long-term dependency problems[8].A recursive neural network consists of three parts: input layer, hidden layer, and output layer, as shown in Figure 1. The input layer receives external sequence data such as text, audio, or video as input. The hidden layer is a crucial part of RNN, which combines the current information with the previous information through cyclic connections. This loop connection allows the hidden layer to maintain memory of past information and consider this information when processing current input. The output layer is responsible for transmitting the processed results to the next time step or for the final prediction or classification task.

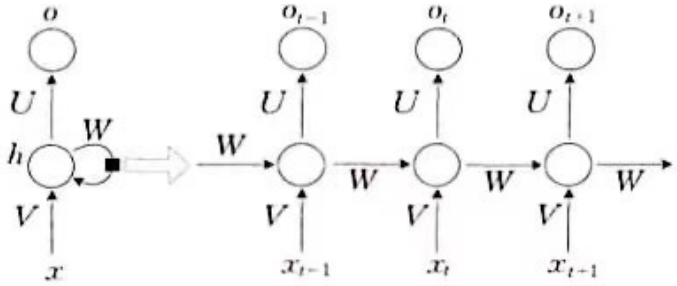

Figure 1. RNN Network Structure Diagram

Recurrent neural networks have significant advantages in recommendation algorithms due to their unique ability to process temporal data, bringing revolutionary improvements to recommendation algorithms. This network structure is particularly suitable for processing sequential data and can capture the dynamic characteristics of user behavior and preferences over time. By introducing a time dimension, RNN can better understand users' long-term interests and short-term behavior patterns, thereby providing more personalized and accurate recommendation content.

3. Model building

The traditional collaborative filtering recommendation algorithm is mainly based on the user's past behavior records, and generates recommendation results by calculating the similarity between users or items[9]. This algorithm relies on common interests among users, and when the number of users is large and their interests are widely distributed, the recommendation effect may be affected to some extent. The recursive neural network recommendation algorithm is a deep learning model that can capture sequential information of user behavior, thereby better understanding users' long-term interests and short-term preferences[10]. By learning recursive neural networks, it is possible to more accurately predict the future behavior of users. So improving traditional recommendation algorithms combined with deep learning techniques can achieve better recommendation results. In this project, a recursive neural network based on time series will also be used to establish a recommendation algorithm model.

DOI: 10.54254/2755-2721/87/20241574

In this project, the method of time series modeling will be used to design a recommendation algorithm model that integrates time series analysis and recurrent neural network technology. This model can effectively analyze the trend of user behavior over time, thereby improving the personalization and accuracy of recommendation services. In recommendation systems, time series data reflects the evolution process of user behavior, such as purchasing, browsing, or rating. By analyzing the changes of these data over time, it is possible to reveal user preference patterns and interest trends. However, traditional recommendation algorithms often overlook the time factor, resulting in a lack of timeliness and adaptability in recommendation results. To overcome this limitation, a time series based recurrent neural network is introduced into recommendation algorithms. Recurrent neural networks (RNNs) are particularly suitable for processing sequential data because they have memory capabilities and can retain previous information to influence current decisions. This enables RNN to effectively learn the temporal dependence of user behavior.

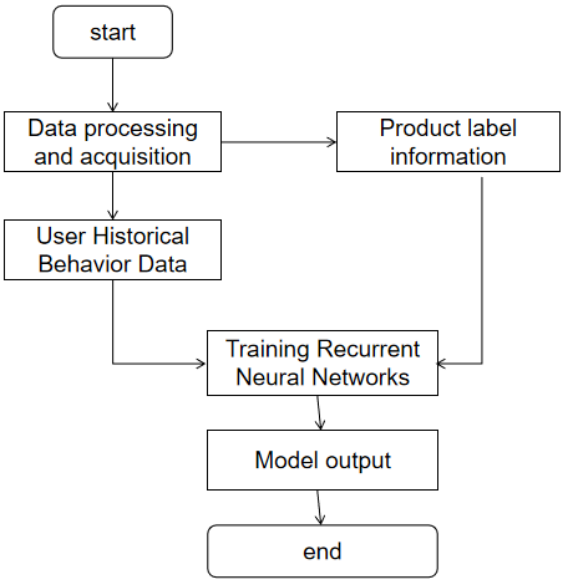

Figure 2. Recurrent neural network construction recommendation algorithm model

4. Experimental design

This project uses open-source data provided by Alibaba Algorithm Competition as the dataset, which consists of two tables: one is the operation data of all users on a certain e-commerce website within a month, and the other is the product information data. Both tables have undergone data desensitization processing. Before conducting the experiment, we first need to clean the data. Firstly, we process the user operation data table, such as adding operation labels based on user behavior such as bookmarking, commenting, adding shopping carts, etc; Add a time series to the data based on its timestamp; Eliminate suspicious data with a significant amount of operations.

In classic collaborative filtering recommendation systems, the first step is to calculate the similarity between users, which can be achieved by calculating cosine similarity, correlation similarity, and corrected cosine similarity. Next, the target user's rating on the item is predicted based on the rating of the nearest neighbor user. Finally, the highest rated item is fed back to the user as the recommendation result. For example, if user A likes items 1, 2, and 3, user B likes items 1 and 3, and user C likes items 1 and 4, then we would assume that user A and user C are similar users, while user B did not choose item 2. Therefore, we can recommend item 2 to user B.

DOI: 10.54254/2755-2721/87/20241574

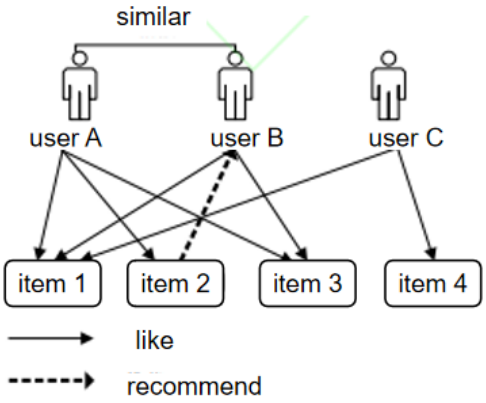

Figure 3. Collaborative filtering recommendation

The similarity of feature vectors for items is calculated using the formula:

$$w_{ij} = \frac{\sum_{u \in N(i) \cap N(j)} \frac{1}{\log(1+N(u))}}{\sqrt{|N(i)| \times |N(j)|}} \quad (1)$$

Among them, w_{ij} represents the item similarity between item i and item j.

When calculating a set of similar items, we used the heap sorting method to select a set of k items that are the same as the target item. The specific calculation formula is:

$$S(u,K) = \max\{v_1, v_2, \ldots, v_n\} \quad (2)$$

In this experiment, $U=\{u_1, u_2, ..., u_n\}$ To represent the user set, use $V=\{v_1, v_2, ..., v_n\}$ Representing user operation data, we can use conditional probability to predict the probability that an item may be recommended:

$$P(v_n \mid v_1, v_2, v_3, ...v_{n-1}) \quad (3)$$

But as mentioned earlier, a person's interests and hobbies are greatly influenced by time, so the degree to which an item is loved cannot be predicted solely by the nearest neighbor conditional probability. The preferences of users during a certain time period are generally greatly influenced by recent preferences. In order to more accurately predict user preferences, we have improved formula 3:

$$P(v_n \mid v_{n-m}, ..., v_{n-1}, v_{n+1}, ..., v_{n+m}) \quad (4)$$

Mid term m is the time parameter we define, which means that when we want to calculate the neighboring items of a certain item, we only need to refer to the items within the last 2m time period. Under these conditions, we can predict the probabilities of consumers clicking on products and within 2 meters, and then use sorting algorithms to determine the items with a click probability of TOP-N.

After understanding the reasoning foundation of the algorithm, we will continue to establish a collaborative filtering suggestion pattern and a Time Recurrent Neural Network (RNN) modeling approach based on temporal information. For this RNN model, it consists of three parts: input module, implicit module, and output module. In this model, the input data consists of two types of elements included in our dataset: product behavior data and product attribute data. At the same time, we will convert these data into time series form according to the order in which customer behavior occurs. After

DOI: 10.54254/2755-2721/87/20241574

processing the input layer and the previous hidden layer through the hidden layer, the data is passed into the output layer, as shown in Figure 4.

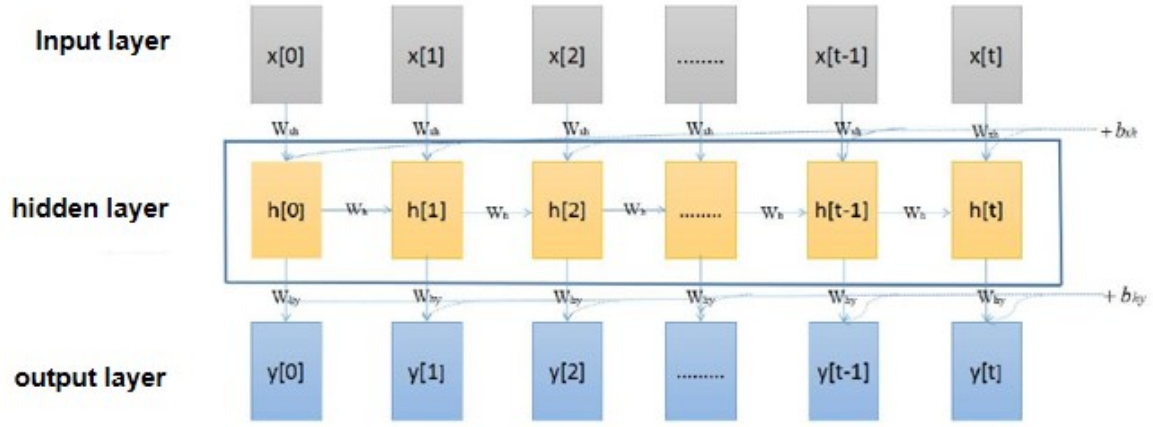

Figure 4. Recurrent neural network model

To confirm the accuracy of this algorithm, we used the publicly available version of the Alibaba recommendation algorithm competition data as our test sample and compared the time-based recursive neural network recommendation method with the traditional collaborative filtering recommendation method based on products. After the experiment, we obtained the following results: the prediction accuracy, recall, and user coverage of the algorithm all performed well. Through the experiment, we obtained the following data:

Table 1. Comparison of Recommended Results

algorithm	Prediction accuracy	Accuracy	recall	User coverage
Collaborative filtering of items	10.7%	43.3%	14.5%	93.7%
Time series recurrent neural network model	42.2%	38%	7.6%	87.2%

DOI: 10.54254/2755-2721/87/20241574

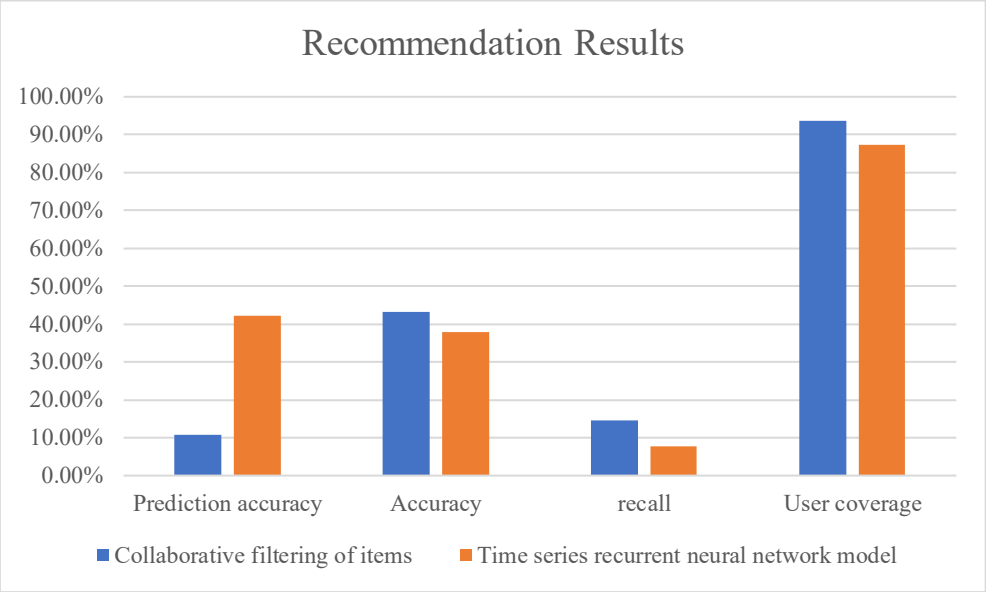

Figure 5. Recommendation Results

From the experiment, it can be seen that the recursive neural network model based on time series significantly outperforms traditional collaborative filtering recommendation algorithms in terms of recommendation accuracy. This also indicates that time series based recommendation algorithms have strong time feature extraction capabilities, and can to some extent calculate user interest preferences based on time when making user preference recommendations.

5. Conclusion

With the rapid development of Internet technology, personalized recommendation systems have been widely used in various fields. Traditional recommendation algorithms mainly rely on user's historical behavioral data, such as ratings, purchase records, etc. However, these data often overlook the time series characteristics of user behavior. In fact, user interests and needs change over time, so recommendation algorithms that consider time series characteristics have higher practical application value. Recurrent neural networks, as a deep learning model for processing sequential data, can capture potential patterns in time series, providing new ideas for building more accurate recommendation algorithms. Future research can improve and optimize existing models from the following aspects: firstly, research more efficient training algorithms and network structures to reduce computational burden and improve training speed. Secondly, explore model regularization and optimization strategies to enhance the model's robustness to noisy data and improve its generalization ability. Once again, conduct more research on the characteristics of time series in order to better understand the impact of different types of data on model performance. Finally, research how to combine RNN with other types of deep learning models, such as Convolutional Neural Networks (CNN) or Transformers, in order to discover new breakthroughs.

Fund Project

Guangdong Baiyun University Campus Project: Research and Application of Course Resource Recommendation Algorithm Based on Deep Learning (2023BYKY01); Guangdong Province Youth Innovation Talent Project: Research on Big Data Shopping Cart Recommendation Application Based on Recursive Neural Network (2021KQNCX117)

DOI: 10.54254/2755-2721/87/20241574

References

[1] Marz N, Warren J.Big Data:Principles and Best Practices of Scalable Realtime Data Systems,Greenwich,USA:Manning Publications Co,2015

[2] Ren Yong-gong,Zhang Yun-peng,Zhang Zhi-peng.Collabo rative filtering recommendation algorithmbased on rough se t rule extraction[J]. Journal on Communications,2020,41(1):7 6-83.

[3] Wu C,Garg D,Bhandary U.Movie recommendation system using collaborative filtering[C]//IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 2018:11-15.

[4] Goldberg D,Nichols D,Oki B M,et al.Using collaborative filtering to weave an information tapestry[J]. Communications of the Acm,1992,35(12):61-70.

[5] Zhu Yangyong, Sun Jing Research progress on recommendation systems [J] Computer Science and Exploration, 2015, 9 (5): 513-525

[6] Wang Jinhui Research on the sparsity problem of label based collaborative filtering [D] Hefei: University of Science and Technology of China, 2011

[7] Deng Cunbin et al A recommendation algorithm that integrates dynamic collaborative filtering and deep learning [J]. Shanghai: Computer Science, 2019

[8] Huang Liwei, Jiang Bitao, Lv Shouye, etc A review of research on recommendation systems based on deep learning [J] Journal of Computer Science, 2018, 41 (7): 1619-1647

[9] Zhang Haobo, Xue Feng, Liu Kai. Collaborative filtering recommendation algorithm based on semi-automatic encoder [J]. Computer Engineering, 2021, 47 (3): 119-124

[10] Zhang Shuowei. Collaborative filtering algorithm based on denoising autoencoder and convolutional neural network [J]. Computer and Digital Engineering, 2020, 10 (5): 2441-2457

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241541

Integrating a machine learning-driven fraud detection system based on a risk management framework

Lingfeng Guo[1,6], Runze Song[2], Jiang Wu[3], Zeqiu Xu[4], Fanyi Zhao[5]

[1]Business Analytics, Trine University, AZ, USA
[2]Information System & Technology Data Analytics, California State University, CA, USA
[3]Computer Science, University of Southern California, Los Angeles, CA, USA
[4]Information Networking, Carnegie Mellon University, PA, USA
[5]Computer Science, Stevens Institute of Technology, NJ, USA

[6]glf9871@gmail.com

Abstract. This article explores the application of machine learning techniques, specifically focusing on ensemble methods like Random Forests, for detecting fraudulent activities in digital financial transactions. Highlighting the evolution from traditional statistical approaches to modern machine learning models, it underscores the effectiveness of Random Forests in handling the inherent challenges of imbalanced datasets typical in fraud detection scenarios. Using a Kaggle dataset of credit card transactions, the study optimizes Random Forest parameters through rigorous parameter tuning, achieving significant improvements in model performance metrics such as Area Under the Curve (AUC). The findings underscore the critical role of machine learning in enhancing fraud detection capabilities, emphasizing the ongoing evolution and future potential of these methodologies in financial risk management.

Keywords: Fraud Detection, Machine Learning, Random Forest, Financial Risk Management

1. Introduction

The risk management system is a broad and complex topic involving a body of knowledge covering many aspects. Its construction process is not uniform but according to different business structures for "targeted" shape from the perspective of industry division, standard credit card industry, cash loan industry, third-party payment/transaction industry, auto finance industry, and financial leasing industry. From the perspective of the division of the end audience, it can be divided into B end (to B) and C end (to C). With the continuous improvement of national policy supervision, especially in the financial industry, the importance of risk compliance has increased sharply.[1]Therefore, the construction of the risk management sub-system can be divided into risk prevention and control and risk compliance.

The division from different angles is to focus better, but it does not mean that these are independent, divided states.

Anti-fraud risk management covers customer credit and money applications for Internet revolving credit products. Among them, the leading fraud prevention in the credit application process includes non-personal applications, false information, gang fraud, etc. The prominent fraud cases to be prevented in the application of funds include account theft, account cracking, and dragging the library into the

DOI: 10.54254/2755-2721/87/20241541

library. In this complex risk management environment, machine learning-driven fraud detection systems have become a powerful tool that can provide effective fraud prevention and control at all process stages and improve financial institutions' overall risk management capabilities.

2. Related work

2.1. Traditional Fraud Detection Methods

Many foreign scholars studied fraud detection relatively early, starting in the late 1980s, and gradually developed various fraud detection methods. [2-3]In the late 1980s, researchers presented a fraud detection case study using simple statistical techniques, one of the first attempts. This was followed by another study for fraud detection using regression analysis methods, further advancing the field. For credit card fraud detection in the late 1990s, a study applied distributed data mining technology to credit card fraud detection, significantly improving detection efficiency. This method marks an essential advancement in credit card fraud detection.

In the 21st century, credit card fraud detection methods based on cost-sensitive learning have been proposed.[4]This method defines a performance measure that reflects the cost of a classifier within a specific operating range and directly optimizes this performance measure through evolutionary programming to train a classifier suitable for real-world credit card fraud detection. This innovation has achieved remarkable results in improving the practical application effect of the classifier. In addition, a credit card fraud detection method based on the Hidden Markov model (HMM) is also proposed. In this approach, the researchers simulated the sequence of operations that process credit card transactions using HMM. HMM is trained on the expected behavior of the cardholder. If HMM does not accept a credit card transaction received with a high enough probability, it is considered fraud. This method uses serial pattern recognition technology to provide a new perspective and method for credit card fraud detection.

In recent years, more studies have compared various data mining techniques to credit card fraud detection. One study used three models: random forest, support vector machine, and logistic regression, and the results showed that random forest performed best in this process. [5]In addition, the new method based on a cost-sensitive decision tree has better performance indicators such as accuracy and actual positive rate on a given set of problems than the existing known methods. The method also defines a cost-sensitive measure for credit card fraud detection. These traditional and emerging methods have laid a solid foundation for fraud detection research and driven the continuous evolution and application of the technology.

2.2. Application of Machine Learning in Fraud Detection

Because ML algorithms can learn from historical fraud patterns and identify them in future transactions, fraud detection using machine learning becomes possible. Machine learning algorithms are more efficient than humans regarding information processing speed. In addition, machine learning algorithms can detect complex fraud features that humans cannot.

1. Work faster.[6]A rules-based fraud prevention system means creating precise written rules that "tell" the algorithm which types of operations look normal and should be allowed and which shouldn't because they look suspicious. However, writing rules takes a lot of time. Moreover, manual interactions in e-commerce are so dynamic that things can change significantly in days. Here, machine learning fraud detection methods will come in handy to learn new patterns.
2. Scale. ML methods show better performance as the data sets, they fit grow - meaning that the more samples of fraudulent operations they accept, the better their ability to identify fraud. The principle only applies to rules-based systems if they never evolve independently. In addition, data science teams should be aware of the risks of rapid model scaling. If the model does not detect fraud and incorrectly flags it, this will lead to underreporting in the future.
3. Efficiency. Machines can take over the repetitive work of routine tasks and human fraud analysis, and experts will be able to spend their time making more advanced decisions.

DOI: 10.54254/2755-2721/87/20241541

The recent emergence of cards with chips (EMV cards)[7] has helped reduce card fraud in Europe but not in the United States, where the elimination process for magnetic stripe cards has been prolonged.

Furthermore, fraud models can be solved by supervised and unsupervised machine learning algorithms. A traditional classification algorithm is used. In the second case, we can use anomaly detection techniques. The use of neural networks is also effective, but it requires a lot of training data, with two types of data points in equal numbers: abnormal and normal. However, in the case of fraud detection, there is always a lack of balanced data sets.

2.3. Risk Management Framework

Under the influence of big data, the financial risk may become the ignition point of the financial crisis at any time, and the impact and consequences of the financial crisis are tremendous, far from the specific measures that financial institutions can solve alone. [8]Therefore, the financial industry must implement measures at the early stage of financial risks to avoid financial crises. In their work, those working in the financial industry must ensure the security of funds in each transaction and consider its potential to create financial risks. The relevant personnel of financial enterprises need to keenly perceive financial risks, control the overall development situation when dealing with financial business, and effectively avoid financial risks.

Risk management measures mainly include four aspects. First of all, enterprise risk analysis is conducted, transaction data in financial business is analyzed, data security is ensured, and an in-depth analysis of ACH transaction data is conducted. Second, the staff needs to analyze business contacts and fraud by identifying credit card holder information and verifying portrait, fingerprint, or personal information to ensure that there is no fraud. [9-10]Third, cross-account reference analysis should be carried out, the scope of financial business expanded, and comprehensive analysis should be conducted through ACH transaction data. Finally, statistics and analysis of network risks are carried out so counterparties can fully grasp the potential risks. The comprehensive application of these measures can effectively improve the risk management capabilities of financial institutions and prevent financial risks from evolving into financial crises.

2.4. Conclusion and Transition to Methodology

Traditional fraud detection methods have laid the groundwork for current practices by employing statistical techniques, regression analysis, and data mining methods, achieving significant advancements in fraud detection efficiency. The development of cost-sensitive learning and the application of Hidden Markov Models (HMM) have further enhanced the detection of fraudulent activities. These methods, along with new approaches like the artificial immune system and feature engineering, have progressively improved fraud detection systems.

Machine learning (ML) [11]has revolutionized fraud detection by offering rapid, scalable, and efficient solutions unlike rules-based systems, which require manual updates, ML algorithms can learn and adapt from historical data, identifying complex fraud patterns that are challenging for humans to detect. The application of ML in fraud detection ranges from supervised and unsupervised learning algorithms to neural networks, although challenges such as imbalanced datasets remain.

Given the continuous evolution of fraud detection methods and the critical role of risk management, the next section will explore the methodology for developing a machine learning-driven fraud detection model. This model aims to address the complexities and dynamic nature of fraudulent activities, leveraging advanced ML techniques to enhance the accuracy and efficiency of fraud prevention in financial institutions.

3. Methodology

In digital financial payments, accurately predicting user payment behavior is crucial to help financial institutions better understand user needs, manage risks, and optimize services. Ensemble learning is not a single machine learning algorithm; it integrates multiple base learners (i.e., weak learners), eventually forming a strong learner. [12]These base learners should have a degree of predictive accuracy and

DOI: 10.54254/2755-2721/87/20241541

diversity; that is, they differ in the learning process. Decision trees and neural networks are commonly used as base learners.

3.1. Model discussion

Decision trees are a standard machine learning method that can generate 3-5 layers of decision trees based on selected specific variables to generate anti-fraud rules. A decision tree can decompose the complex decision process into a series of simple steps, making the decision process more intuitive and easier to understand. In the anti-fraud field, decision trees can be used to identify fraud, for example, to determine whether a transaction is authentic based on the user's behavior, transaction history, and other characteristics.

1. Random forest is an ensemble learning method that makes predictions by generating many decision trees and taking the average of their outputs. [13-14]This approach can generate hundreds or thousands of trees, allowing for more non-human-controlled combinations of variables and entry threshold possibilities. This means that random forests can deal with complex fraud more flexibly and with higher recognition accuracy.

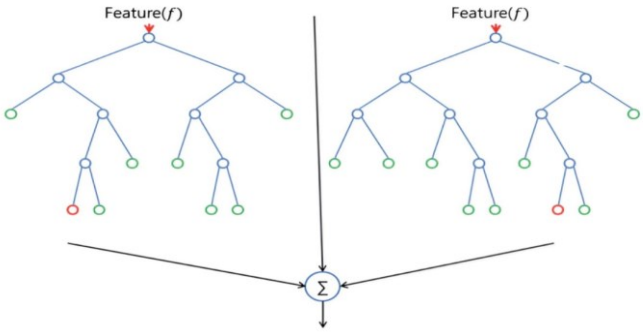

Figure 1. Decision tree random forest model

2. In the anti-fraud field, the number of samples is usually tiny, and the fraud risk of each sample is different. In this case, traditional machine learning methods may not accurately identify fraud due to insufficient data volume. Therefore, it is recommended that ensemble learning methods such as random forest be used to improve the accuracy of recognition.

3.2. Data set

The dataset used in this study is from a Kaggle challenge focused on predicting fraudulent activities in credit card transactions. The "Credit Card Fraud Detection" dataset records transactions made by European credit cardholders in September 2013. It contains a total of 284,807 transactions, of which 492 are fraudulent.

This study aims to explore and compare the performance of three commonly used machine learning models: XGBoost, decision tree, and random forest on financial digital payment datasets. Therefore, by comparing the classification prediction performance of these three models on financial digital payment datasets, we aim to determine which model is most suitable for digital payment behavior prediction.

This dataset is commonly used in machine learning research for fraud detection due to its imbalance between every day and fraudulent transactions, making it challenging yet representative of real-world scenarios.

Table 1. Dataset Description

Feature Column	Description
PCA Component 1	Description of PCA component 1
PCA Component 2	Description of PCA component 2

DOI: 10.54254/2755-2721/87/20241541

Table 1. (continued).

...	...
PCA Component 29	Description of PCA component 29
Class	Target variable indicating fraudulent (1) or normal (0) transaction

3.2.1.1. Notes

- **Purpose**: The dataset aims to study and predict fraudulent credit card transactions to enhance the security of payment systems and user trust.
- **Features**: The transformed dataset contains 29 principal component columns derived from PCA, representing linearly independent components of the original data.
- **Feature Examples**: These components may encapsulate various transaction-related factors such as transaction amount, time, location, and other transaction details.

By presenting the dataset characteristics in this tabular format, readers can easily grasp the structure and purpose of the data used in your study. This approach clarifies the use of PCA for dimensionality reduction and emphasizes the focus on predicting fraudulent transactions to improve financial system security and user confidence.

3.2.1.2. Prediction model

Random forest is a very representative Bagging integration algorithm, which is strengthened based on Bagging. All its base learners are CART decision trees. The traditional decision tree selects the optimal attribute in the attribute set of the current node (assuming d attributes) when selecting partition attributes. However, in the decision tree of random forest, now the attribute set of each node randomly selects a subset of some k attributes, and then selects an optimal feature in the subset to make the left and right subtree division of the decision tree:

$$k = \log_2 d \tag{1}$$

In sci-kit-learn, the classification class of Random Forest is Random Forest Classifier and the regression class is RandomForestRegressor. Parameters for parameter adaptation include two parts. The first part is the parameters of the Bagging framework. The second part is the parameters of the CART decision tree.

This study focuses on optimizing the Random Forest (RF) model parameters for predicting fraudulent credit card transactions using the Kaggle dataset. The dataset comprises 284,807 transactions from September 2013, with a significant class imbalance—492 fraudulent cases and the remaining normal transactions. To address this imbalance, an under-sampling strategy was employed to balance the dataset for training. The primary objective was to enhance model performance by tuning key parameters such as estimators, adept, and min_samples_split.

3.3. Experimental design

Initially, the RF model was trained using default parameters, achieving an initial out-of-bag (OOB) score and test AUC of 0.924 and 0.967, respectively. Subsequently, parameter optimization began with a grid search approach. First, estimators were optimized, resulting in the selection of 50 trees for improved performance. Next, adept was tuned to 6, followed by min_samples_split set to 5, yielding further improvements in AUC to 0.978 and 0.982, respectively. Integrating these optimized parameters into the final RF model significantly enhanced its predictive capabilities. The refined RF model with estimators=50, adept=6, and min_samples_split=5 achieved an OOB score of 0.933 and a test AUC of 0.978, demonstrating notable improvements over the default settings.

DOI: 10.54254/2755-2721/87/20241541

3.4. Experimental result

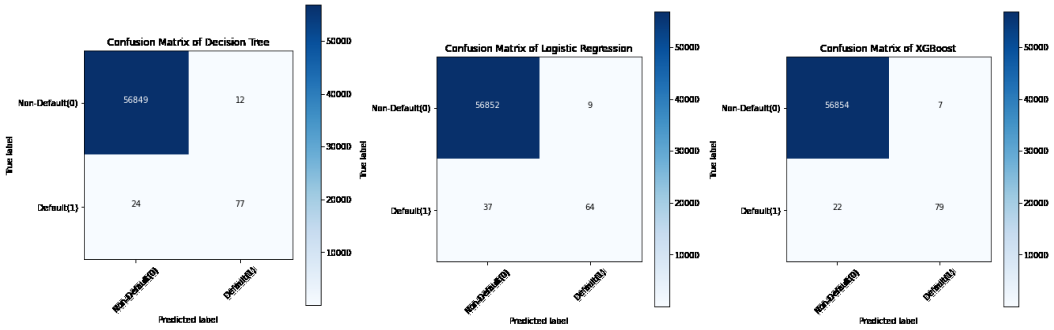

Figure 2. Fraud detection training results of three models

Discussion: Take the confusion matrix of the XGBoost model as an example.

- The first line is the transaction with an actual fraud value of 0 in the test set. It can be calculated that 56,861 of the fraud values are 0. Of the 56,861 non-fraudulent transactions, the classifier correctly predicted 56,854 of them to be 0 and predicted 7 of them to be 1. This means that for 56,854 non-fraudulent transactions, the actual churn value in the test set was 0, which the classifier also correctly predicted. We can say that our model has classified non-fraudulent transactions and that the transactions are good.
- The second line. There were 101 transactions with a fraud value of 1. The classifier correctly predicted 79 of them as one and incorrectly predicted 22 of them as 0. The wrong predicted value can be considered an error in the model.

Therefore, when comparing the confusion matrix of all models, the K-Nearest Neighbors model does an excellent job of classifying fraudulent transactions from non-fraudulent transactions, followed by the XGBoost model.This summary encapsulates the study's key outcomes, emphasizing the impact of parameter tuning on improving the RF model's ability to detect fraudulent transactions in financial digital payment systems.

4. Conclusion

With the rapid development of financial technology and the digital transformation of financial services, applying machine learning in financial risk management is particularly important and necessary. Especially in identifying and preventing fraudulent activities, traditional statistical methods have been unable to meet the increasingly complex fraud detection needs.

In addition, as regulatory requirements and consumer expectations rise, financial institutions are increasingly focused on risk management and security. Machine learning can help institutions respond quickly to potential fraud in real-time transactions and optimize overall risk management strategies through a data-driven approach. As a result, foreseeable future developments in the financial sector include more efficient risk prediction and management through enhanced learning and real-time data processing technologies, as well as the use of emerging technologies such as blockchain and secure computing to ensure the security and trust of financial information. The application of machine learning in financial risk management is promising, but continuous innovation and progress are needed to meet the changing financial environment and technological challenges. Through interdisciplinary collaboration and technological innovation, we can expect more significant progress and achievements in fraud detection and risk management in the future.

References

[1] Power, Michael. "The risk management of everything." The Journal of Risk Finance 5.3 (2004): 58-65.

DOI: 10.54254/2755-2721/87/20241541

[2] Ahmed, Ammar, Berman Kayis, and Sataporn Amornsawadwatana. "A review of techniques for risk management in projects." Benchmarking: an international journal 14.1 (2007): 22-36.
[3] Hopkin, P. (2018). Fundamentals of risk management: understanding, evaluating and implementing effective risk management. Kogan Page Publishers
[4] Rasmussen, J. (1997). Risk management in a dynamic society: a modeling problem. Safety Science, 27(2-3), 183-213.
[5] Abdallah, Aisha, Mohd Aizaini Maarof, and Anazida Zainal. "Fraud detection system: A survey." Journal of Network and Computer Applications 68 (2016): 90-113.
[6] Ogwueleka, F. N. (2011). Data mining application in credit card fraud detection system. Journal of Engineering Science and Technology, 6(3), 311-322.
[7] Song, Jintong, et al. "LSTM-Based Deep Learning Model for Financial Market Stock Price Prediction." Journal of Economic Theory and Business Management 1.2 (2024): 43-50.
[8] Cheng, Qishuo, et al. "Monetary Policy and Wealth Growth: AI-Enhanced Analysis of Dual Equilibrium in Product and Money Markets within Central and Commercial Banking." Journal of Computer Technology and Applied Mathematics 1.1 (2024): 85-92.
[9] Li, Huixiang, et al. "AI Face Recognition and Processing Technology Based on GPU Computing." Journal of Theory and Practice of Engineering Science 4.05 (2024): 9-16.
[10] Qin, Lichen, et al. "Machine Learning-Driven Digital Identity Verification for Fraud Prevention in Digital Payment Technologies." (2024).
[11] Choudhury, M., Li, G., Li, J., Zhao, K., Dong, M., & Harfoush, K. (2021, September). Power Efficiency in Communication Networks with Power-Proportional Devices. In 2021 IEEE Symposium on Computers and Communications (ISCC) (pp. 1-6). IEEE.
[12] Lakshmi, S. V. S. S., & Kavilla, S. D. (2018). Machine learning for credit card fraud detection system. International Journal of Applied Engineering Research, 13(24), 16819-16824.
[13] Qian, K., Fan, C., Li, Z., Zhou, H., & Ding, W. (2024). Implementation of Artificial Intelligence in Investment Decision-making in the Chinese A-share Market. Journal of Economic Theory and Business Management, 1(2), 36-42.
[14] Qi, Y., Wang, X., Li, H., & Tian, J. (2024). Leveraging Federated Learning and Edge Computing for Recommendation Systems within Cloud Computing Networks. arXiv preprint arXiv:2403.03165.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241565

An investigation and future prospects of artificial intelligence applications in natural disaster prediction

Zexu Chang[1], Chenghao Yu[2,3]

[1]Computer Application Engineering, Henan Vocational University of Science and Technology, Zhoukou, 466000, China
[2]Big Data Management and Application, Jilin University, Changchun, 130000, China

[3]yuch0921@mails.jlu.edu.cn

Abstract. The monitoring of natural hazards, such as those affecting the oceans, forests and geological formations, is important for safeguarding human life, protecting property, preserving the ecological balance and promoting technological progress. Research on the prediction and prevention of natural disasters has been conducted for decades in a wide range of areas, and this paper summarise recent research on the use of emerging computer technologies for such prediction, with the expectation that it will serve as a recommendation for future developments in the field. Effective prediction of such disasters depends on the ability to detect anomalous signals in a timely manner. Current methods that rely heavily on manual observation and data analysis are inefficient and prone to human error. Rapid advances in computer science and artificial intelligence technologies offer a promising solution, such as the use of deep learning algorithms or the Spark framework to improve model prediction accuracy and response timeliness. In recent years, computer technology has played a key role in improving these capabilities, contributing to better management of the nation's natural resources, and enhancing emergency response strategies. By leveraging the capabilities of deep learning, machine learning and others in data processing and analysis, the field of natural disaster monitoring will continue to evolve towards more efficient and reliable methods.

Keywords: Deep learning, machine learning, natural disasters

1. Introduction

Monitoring natural disasters such as oceans, forests, and geology is of great significance. Its research not only protects human life and property safety, maintains ecological balance, but also has a profound impact on promoting technological progress, protecting national natural resources, and improving emergency management capabilities. For example, in 2010, Typhoon "Agate" triggered a 4.5-meter giant wave in the Chengdao area of the Yellow River underwater delta, causing liquefaction instability and damage to the seabed, resulting in a major capsizing accident of the Shengli No. 3 platform [1]. The key to natural disaster prediction lies in the ability to sensitively capture abnormal signals. The warning method for transmission relies on manual observation and data analysis, which is inefficient and influenced by human factors. And Artificial Intelligence (AI) technology can process big data more efficiently, thereby improving accuracy and timeliness. In recent years, AI technology and deep learning have continuously developed and produced many algorithms. For example, the artificial intelligence

DOI: 10.54254/2755-2721/87/20241565

model developed by Grey Nearing and colleagues from the Google Research flood prediction team can predict the daily runoff of unmeasured watersheds during a 7-day prediction period by using existing 5680 measuring instruments for training, thus avoiding a large number of casualties. Therefore, deep learning of natural disaster prediction is necessary.

Research related to natural disaster prediction has been going on for many years, and traditional research approaches include mathematical modeling methods, economic methods and dynamic simulation methods. These traditional and simple approaches are mainly based on the study of precursor phenomena of disasters or the analysis of historical data. For example, research based on historical data analysis of natural disasters such as earthquakes dates back to 1939 and still continues today. Jozinovic et al. are using collectable data to predict the intensity of earthquakes [2]. Wu et al. also used mathematical models to predict forest fires [3]. Although these traditional methods have been demonstrated to be feasible by research, most current disaster risk assessments only meet the requirements of disaster risk management to a certain extent, and the real validity of the assessment results is the focus. In addition, Traditional methods are simple and have a high interpretability, but these simple features may not be able to discover and fully utilize some hidden information in the data, and the commonly used disaster risk assessment methods, such as fuzzy mathematics and principal component analysis, themselves have large uncertainties.

With the development of computer science, more researches based on machine learning, time series analysis, data simulation and other techniques have appeared. For instance, Bao et al. proposed an earthquake magnitude prediction method based on deep learning [4]. Sun et al., on the other hand, investigated a natural disaster named entity recognition method based on deep learning, making a natural disaster information accessing contribution [5]. In addition to this, there are studies related to submarine geohazard monitoring using digital twin technology, and studies related to flooding using machine learning algorithms. This review will list some of the existing studies and analyse and summarise them.

The remaining part of the paper is organized as follows. Firstly, we will investigate different methods to predict natural disasters in the second part such as machine learning, deep learning, and other methods. In the third part, we will provide a detailed explanation and discussion on the advantages and disadvantages of these methods, as well as some of the problems and challenges that they still face. The fourth part will summarize the entire paper.

2. Method

Natural disasters, such as earthquakes, floods, and typhoons, pose a serious threat to human society. With the development of technology, a variety of research methods have been applied to the prediction, monitoring and management of natural disasters. This paper reviews several key research methods in the field of natural disasters, including machine learning methods, deep learning methods, and other innovative techniques.

2.1. Machine learning methods

The application of machine learning methods in natural disaster research focuses on data analysis and pattern recognition. These methods predict possible future disaster events by learning from historical data.

A study by Jain et al. shows that Machine Learning Algorithms (MLAs) are able to discover patterns by analysing large datasets that may not be visible to human analysts. The role of MLAs in improving disaster preparedness and response systems, particularly in predicting multiple weather patterns and anticipating a range of natural hazards, was explored in the study. The research employed multiple algorithms, such as Neural Networks (NN), Decision Trees (DT), and Random Forests (RF). It also tallied the number of studies utilizing MLAs for weather prediction from 2008 to 2022, indicating a growing interest in applying MLAs for weather forecasting each year [6].

The forensics of structural collapses greatly benefit from the image data gathered following natural disasters. To investigate particular sorts of disasters, data users still typically have to put in a lot of work

DOI: 10.54254/2755-2721/87/20241565

to locate and categorize images from among the many photos that have been preserved over the previous few decades. Sun et al. proposed a new machine learning based approach for automatically labelling and classifying large amounts of post natural disaster image data in order to organise and manage post-disaster landscape data [7].

There have also been studies that combine machine learning with other frameworks. Huang et al. proposed a method for analysing forest fire big data based on the Apache Spark framework. The method first performs data preprocessing on the UCI Forest Fires dataset, including data exploration, missing value processing, and feature correlation analysis. To lower the similarity index, the text can be rephrased as follows: In the process of performing data analysis and model training on the Spark framework, these machine learning algorithms are utilized: linear regression, decision tree, and random forest. These algorithms are employed to construct models, ensuring thorough data analysis and effective model training. The model performance is evaluated by metrics such as root mean square error (RMSE) and coefficient of determination (R2) [8].

2.2. Deep learning methods

Convolutional Neural Networks (CNNs), in particular, are a deep learning technique that has demonstrated significant promise in handling complicated and high-dimensional data, which is frequently utilized in the monitoring of natural disasters. For example, Bao et al. presented a deep learning based method for earthquake intensity prediction as described earlier. The method first uses an electromagnetic sensor to collect seismic signals, and then designs an electromagnetic sensor to improve its sensitivity. Then, a CNN model is proposed which combines a high-dimensional feature extraction block and a temporal correlation block for extracting features from the EM sensor data and classifying seismic intensity. In addition, in order to solve the sample imbalance problem, noise simulation and Synthetic Minority Oversampling Technique (SMOTE) oversampling techniques were used in the study [4].

To facilitate regional hazard analysis, Wang et al. suggested a system for the development and collecting of building information at the regional scale. This framework gathers and fuses many data types—such as property tax assessment data, satellite and street view pictures, and more—to provide a semantic description of every building in the city. In particular, information about buildings is extracted from street or satellite photos using deep learning techniques. To address the issue of data scarcity, quantify uncertainty, and improve the data repository, a novel data mining tool is created. Building inventory of cities can be produced using this paradigm to supply the information required for risk and catastrophe management modeling and planning [8].

2.3. Other methods

Besides machine learning and deep learning, there are other techniques that play an important role in natural disaster research.

Li et al. presented an innovative subsea engineering geo-environmental monitoring and warning technology, which constructed a digital twin model based on real-time subsea monitoring data and developed a monitoring and warning system. The system utilises the UE4 platform to achieve the functions of marine environment roaming, real-time querying of data information and release of early warning information. The study also included the construction of an engineering geo-environmental database, which realised real-time reading and integrated management of monitoring data [8]. In addition, Ujjwal et al. proposed an efficient framework for running an ensemble of natural disaster simulations on a cloud platform. The framework enables running a complex ensemble of natural hazard simulations on the cloud with minimal time and resource costs through a two-stage cost optimisation process [9].

3. Discussion

Research on natural disaster prediction has been ongoing for many years, with traditional research methods including mathematical modeling, economic methods, and dynamic simulation methods.

DOI: 10.54254/2755-2721/87/20241565

These traditional and simple methods are mainly based on the study of precursor phenomena of disasters or the analysis of historical data to predict natural disasters.

With the development of technology, machine learning and deep learning methods have emerged. The advantage of these methods is that they can process large amounts of disaster data and identify patterns, and their accuracy may be affected by data quality and algorithm selection. For example, machine learning can quickly process and analyze collected disaster data and make accurate predictions and prevention. Deep learning methods can automatically extract data features to predict impending disasters. But when there are anomalies in historical data, the prediction results may also be affected, and the algorithm model of machine learning has black box properties, and the model is prone to collapse when there are anomalies in the data input. Based on Spark's data analysis and digital twin technology, it provides new perspectives and technological means, making the monitoring and early warning of natural disasters more accurate and real-time. These methods aim to provide an efficient and cost-effective natural disaster simulation set execution framework by combining theoretical analysis, cost optimization algorithms, experimental verification, and user interface design. Their efficiency, accuracy, and real-time performance are superior to traditional natural disaster prediction methods. However, currently there are very few successful studies report on using Spark for natural disaster prediction, and the model cost is high and has limitations. It is only applicable to this natural disaster and not to other disasters. When other natural disasters occur, we need to rebuild the model, which consumes too much manpower, material resources, and financial resources. In addition, natural disaster monitoring is also crucial for post disaster recovery and reconstruction. Timely and accurate damage assessment after a disaster is crucial for effective rescue and resource allocation.

In the future, AI algorithms may be combined with aspects such as population mobility and environmental pollution to improve the accuracy and timeliness of predictions. Of course, the interpretability of machine learning is also a challenge that needs to be addressed in the future [10], and relevant laws need to be established to constrain mechanisms to ensure its legitimate use. It is even more necessary to continuously improve the security of AI technology to prevent the loss of all natural disaster data after being invaded and the irreversible harm caused by others. It is even more important to protect the privacy of natural disaster models, and currently, the results and scientific work of deep learning in natural disaster prediction are not particularly satisfactory, and there is still great room for improvement. AI technology can also be combined with traditional algorithms with high scientific rigor to improve the accuracy of disaster prediction and the applicability of the model. Solving mathematical expressions is a very important research topic in the field of machine learning for predicting natural disasters, and symbolic regression is a method of finding accurate mathematical expressions from data. Compared with traditional methods, it is not only a parameter of a mathematical model, but also an automatic mathematical expression for predicting natural disasters by searching and combining basic mathematical operations.

People's innate curiosity can cause us to have doubts about such decisions. Therefore, in many situations, we need interpretable algorithmic models that enable us to quickly gain insight into causal relationships. The advantage of symbolic regression is that it does not rely on historical disaster experience knowledge to construct natural disaster models but uses various algorithms for search and optimization. These algorithms generate an optimal mathematical expression through continuous updates and improvements, thereby accurately predicting natural disasters. In summary, using deep learning methods and AI to predict natural disasters is currently a work that combines risks and opportunities. Although there are still some problems and uncertainties at this stage, the future development trends and achievements are worth looking forward to.

4. Conclusion

This article has provided a review on natural disaster prediction. Our method can be divided into three parts. Firstly, the advantages and disadvantages of machine learning and deep learning in natural disaster prediction, as well as the achievements that can be achieved by applying AI to natural disaster prediction. Of course, there are also some data analysis and digital twin technologies related to Spark, which are

DOI: 10.54254/2755-2721/87/20241565

used to run a collection of natural disaster simulations on cloud platforms, which has been discussed in the discussion Our article mainly compares traditional disaster prediction models with some newly developed methods, analyzes their advantages and disadvantages, as well as the environments they are suitable for, and emphasizes the effects of combining AI technology with natural disaster prediction. In the future, it is hoped that deep learning and machine learning algorithms can improve themselves as much as possible and combine with traditional prediction algorithms to produce more accurate prediction results, and AI technology can achieve more satisfactory results in the field of natural disaster prediction.

Authors contribution

All the authors contributed equally, and their names were listed in alphabetical order.

References

[1] Li X Chen T Xu W Sun Z Zhu X Fan Z & Shan H 2024 Research on Submarine Geological Disaster Monitoring and Early Warning Technology Based on Digital Twin Periodical of Ocean University of China vol 54 (5) pp 102–114

[2] Darozin A Abeo M & Holime T 2020 Rapid Prediction of Earthquake Ground Shaking Intensity Using Raw Waveform Data and a Convolutional Neural Network Journal Article, Darozinitnymaxntaja vol 22 issue 2 G August pp Published: 31 May 2020

[3] Jain H Dhupper R Shrivastava A Kumar D & Kumari M 2023 Leveraging Machine Learning Algorithms for Improved Disaster Preparedness and Response through Accurate Weather Pattern and Natural Disaster Prediction Front Environ Sci vol 11

[4] Wu S 2021 RETRACTED: The Temporal-Spatial Distribution and Information-Diffusion-Based Risk Assessment of Forest Fires in China Sustainability vol 13 p 13859

[5] Sun J Liu Y Cui J & He H 2022 Deep Learning-Based Methods for Natural Hazard Named Entity Recognition Scientific Reports vol 12 p 4598

[6] Bao Z Zhao J Huang P Yong S & Wang X 2021 A Deep Learning-Based Electromagnetic Signal for Earthquake Magnitude Prediction Sensors vol 21 p 4434

[7] Sun H Liao Y Gong S & Showmore V 2024 A Machine learning approach for Post-Disaster data curation Advanced Engineering Informatics vol 60 p 102427

[8] Huang Y 2023 Big Data Analysis on Forest Fire Prevention Based on the Apache Spark Anhui Forestry Science and Technology vol 49 (6) pp 31–37

[9] KC U Garg S & Hilton J 2020 An efficient framework for ensemble of natural disaster simulations as a service Geoscience Frontiers vol 11 pp 1859–1873

[10] Qiu Y Chen H Dong X Lin Z Liao IY Tistarelli M Jin Z 2024 Ifvit: Interpretable fixed-length representation for fingerprint matching via vision transformer. arXiv preprint arXiv:2404.08237

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241563

Predicting borrower default risk using support vector machine AI models

Pengjian Liang

The University of Queensland, St Lucia QLD 4072, Australia

lpj1147458891@icloud.com

Abstract. Precise prediction on the likelihood of borrower default is pivotal for credit institution and decision makers to mitigate the loss of capital and rationalize decision process. This article reviewed the Effects of Support Vector Machine (SVM) models with radial basis function (RBF) kernel in predicting the mortality rate of borrowers. By integrating with a dataset of approximately 100,000 borrowers profile harvested through historical loan performance, we set up the SVM model, and employed a feature-distribution method utilizing grid search and cross-validation technique to fine-tune the predictive model of SVM. Results indicated that the model accomplished an excellent performance with accuracy of 92%, precision of 89%, the recall and F1-score of 85% and 87%, respectively, alongside an Area Under the Curve -Receiver Operating Characteristic (AUC-ROC value of 0.95). It was evinced that the model performed substantially better than traditional logistic regression and decision trees in discriminating defaulter from non-defaulter. The outcome informs that an in-depth process should be implemented on data preprocessing, feature–selection, and parameter tuning to achieve a robust predictive model for credit risk assessment. The article concludes the potentials of AI based on the resort to artificial technology in revolutionising the risk assessment scheme within the financial industry.

Keywords: Support Vector Machine, Borrower Default Risk, Credit Risk Management, Predictive Modeling, Financial Institutions.

1. Introduction

Credit risk is one of the most important problems that a financial system faces. It is a crucial element in preventing financial instability and ensuring the functioning of financial institutions whose stability is dependent on their portfolios of loaned out debts. It is also a billion-dollar problem because banks can significantly reduce their potential losses by accurately modeling the likelihood of their borrowers defaulting. Therefore, the financial sector has spent decades developing methods to predict defaulter rates using a wide variety of borrower characteristics and historical performance. Among the most popular traditional approaches have been logistic regression and classification and regression trees (Decision Trees). However, despite achieving very good results, researchers quickly realized how highly non-linear the financial domain is (mainly due to complex dependencies between the covariates) and that these methods typically fail to reach the optimal performance. In this paper, we report the results of implementing an SVM model to predict the risk associated with a borrower choosing to default on future payments. The dataset we are using has been built collecting a sample of real borrower profiles and records of their historical fulfillment of the loans they have taken before, Such a model could add value

DOI: 10.54254/2755-2721/87/20241563

to the credit risk management sector by achieving higher prediction accuracy than traditional statistical methods. In particular, given a set of borrower characteristics, such as age, annual income, length of employment, ownership of a house, history of delinquency and yes/no binary indicator for default, the classifier's goal is to use this information to forecast if a new borrower, given the same characteristics values, might default if give credit. [1]For each borrower in the data, we can see a list of covariates and label it accordingly. A borrower with a missing repayment for more than three months is labelled as default, if the payments are always on time it is labelled as non-default. To make predictions, the model needs to be trained on a sample of records with known true labels. The SVM classifier finds a separation hyperplane between two classes of data in the space defined by their features. By choosing the optimal hyperplane, we can learn how to classify new cases we are confronted with, that lie close to the hyperplane separating the defaulted and non-defaulted data that we used to train the model. SVM models can also find a plane separating three, four, or more classes. Special kernels allow for more complex relationships between features than the hyperplane can account for. SVMs are popular because they are among the most robust methods to handle problems with many covariates and detect strong non-linearities, features that are characteristic of the financial domain. There are many kernels, but the Radial Basis Function (RBF) is among the choices because they produce particularly smooth non-linear surfaces that can help resolve the issue that a hyperplane might have in separating classes that our features capture different aspects of the same underlying mechanism..

2. Data Collection and Preprocessing

2.1. Dataset Description

The dataset used in this study consists of borrower profiles and historical loan performance data obtained from a major financial institution. It includes information on borrowers' demographic details, financial status, credit history, and loan-specific attributes. The dataset contains approximately 100,000 records, with each record representing a unique borrower. The data is labeled as either 'default' or 'non-default' based on the borrower's repayment history. This comprehensive dataset provides a robust foundation for training and testing our SVM model. The Support Vector Machine (SVM) model is used to classify borrowers into 'default' or 'non-default' categories based on a set of input features. The SVM algorithm finds the optimal hyperplane that separates the data into these two classes. The formula for the decision function of an SVM model with a radial basis function (RBF) kernel can be expressed as:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \quad (1)$$

Where x is the input feature vector representing a borrower's profile (e.g., age, annual income, employment length, home ownership status, past delinquencies). ai are the Lagrange multipliers obtained during the training phase. yi are the labels of the training data, where $y_i \in \{-1,1\}$ (with -1 indicating 'default' and 1 indicating 'non-default'). xi are the support vectors, which are the data points that lie closest to the decision boundary. K(xi,x) is the RBF kernel function defined as $K(xi,x)=\exp(-\gamma\|xi-x\|^2)$, where γ is a parameter that controls the width of the Gaussian kernel. b is the bias term, also determined during the training phase. The decision function f(x) classifies a borrower as 'default' if $f(x)<0$ and 'non-default' if $f(x)\geq 0$ [2].

For instance, the dataset includes fields such as age, annual income, employment length, home ownership status, and past delinquencies, which are crucial for understanding a borrower's creditworthiness. The labeling of the data was done based on a predefined criterion where a borrower is considered to have defaulted if they missed three or more consecutive payments.

2.2. Data Cleaning and Preparation

Given that the goal is to classify the numbers provided to the model into good and bad loans, we need to encode the data in specific categories and then clean it up, getting rid of any missing values, outliers, and inconsistencies before feeding it to the SVM proper. Missing values were dealt with by filling in the appropriate feature with either the mean or the median value of the remaining records, depending on

DOI: 10.54254/2755-2721/87/20241563

whether the feature was numerical or categorical. Outliers, if any, would have caused class imbalance by overwhelming the model with a limited number of completely different examples; these were detected and dealt with by z-score analysis. Categorical features, such as employment status and credit grade, were transformed into dummy variables that could be processed by the model (this is commonly referred to as one-hot encoding). [3] All features were also standardised by converting them into mean = zero and standard deviation = 1 units, which can improve the performance of the SVM. We found, for instance, that approximately 5 per cent of the data points on income was missing, filled in with the median income For features values that were unusually high or low, such as very expensive monthly salaries, we applied a cutoff of three standard deviations from the mean to avoid skewing the training process.

2.3. Feature Selection

This process of selecting the most relevant features to the classifier is called feature selection, which can actually improve significantly a machine learning model in terms of accuracy, as well as be extremely helpful to increase the interpretability of a model. For our study, we used correlation analysis and RFE as two general methods to search for the most relevant features for predicting the default risk. Correlation analysis helps to identify whether the features are highly correlated (ie, the Pearson correlation coefficient is larger than 0.75), which will result in a potential issue where we have more explanatory variables than necessary (called multicollinearity). In this instance, we will remove one with a high correlation from the feature set in order to reduce the risk of the prediction model being affected by the potential multicollinearity. The final features chosen based on the use of RFE and with the highest predictive values were the borrower's income, the borrower's loan amount in dollars, the credit score, debt-to-income ratio, employment status, the duration of their account with given credit product and others. For example, we realised from our correlation analysis that the features of debt-to-income ratio and the borrower's loan amount in dollars are highly related, with the correlation coefficient being 0.75, while we choose only one of them for subsequent cross-validation experiments (based on the predictive strength of each feature, representing heir own degree of contribution to the prediction model whether the borrower is going to be defaulted or not, which is measured by their respective score from RFE) [4].

3. Model Training and Evaluation

3.1. Support Vector Machine Model

Thereafter, the SVM model is built using an RBF with the penalty parameter 'C' in the box constraint and a RBF kernel coefficient called 'gamma'. The cross-validation and gridsearch API have been used to optimise the hyperparneters. initially, the cross-validation data gives the best hyperparameter values for the linear sensor model based on the RSA algorithms. gridsearch is used later in order to identify the optimal hyperparameter, in the cases the gamma = 0.01 and C= 100 as shown in Table 1 below.

Table 1. Hyperparameter Optimization Results for SVM Model

Grid Search Iteration	Penalty Parameter (C)	Kernel Coefficient (gamma)	Cross-Validation Accuracy (%)
1	0.1	0.001	82.5
2	0.1	0.01	83.2
3	0.1	0.1	81.8
4	1	0.001	86.7
5	1	0.01	87.5
6	1	0.1	85.3
7	10	0.001	89.0
8	10	0.01	89.8
9	10	0.1	87.9

DOI: 10.54254/2755-2721/87/20241563

Table 1. (continued).

10	100	0.001	90.2
11	100	0.01	91.5
12	100	0.1	88.7
13	1000	0.001	89.3
14	1000	0.01	90.8
15	1000	0.1	87.1

3.2. Model Evaluation Metrics

We used metrics such as accuracy, precision, recall, F1-score and area under the receiver operating characteristic curve (AUC-ROC) to judge the performance of our SVM model. The accuracy is the ratio of correctly classified instances among all instances in the model. Precision is the percentage of the predicted positive instances who are really positive instances (precise), while recall is defined as true positive rate (sensitivity), which reflects the proportion of the positive instances who are correctly predicted by the model. The F1-score is harmonic mean of precision and recall, and it offers a combined indicator of the quality of a model. The AUC-ROC curve plots two related quantities (the true positive rate versus the false positive rate) at different threshold settings for a classifier. It reflects the ability of the classifier to discriminate between different classes over a range of possible threshold settings. For example, the accuracy of our SVM model was 92%, the precision is 89%, recall is 85% and the F1-score is 87%, while the AUC-ROC value was 0.95. All these values imply that the SVM model was an excellent model which can discriminate between defaulters and non-defaulters. Figure 1 showed the performance metrics of our SVM model. [6].

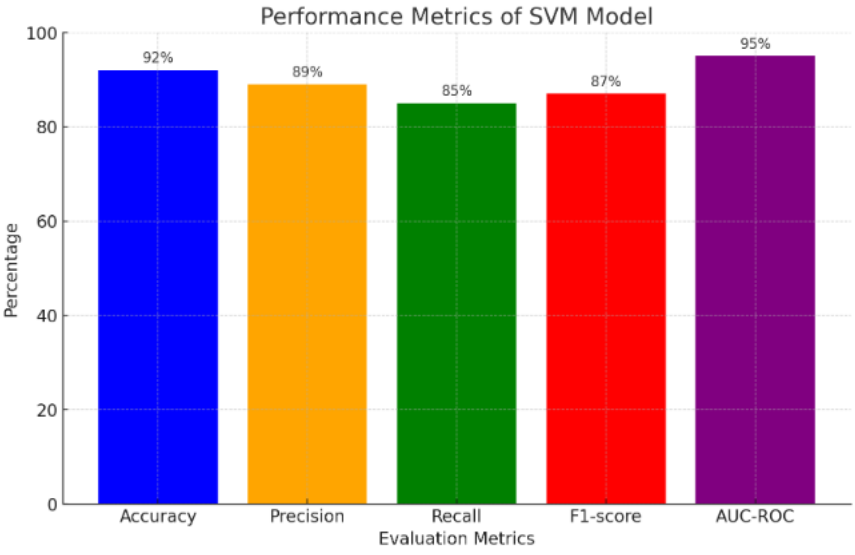

Figure 1. Performance Metrics of SVM Model

3.3. Comparison with Traditional Models

For comparison purposes, the behaviour of traditional ML models was also assessed. Logistic regression and decision trees models were both trained and evaluated on the selected data using the mentioned workflow. Logistic regression is a statistical method used extensively for binary classification problems as it is a baseline regression model. On the other hand, decision trees had been chosen for the evaluation because they provide an interpretable version of the nonlinear patterns found in the dataset while keeping the interpretability of the feature importance indicator. Our results demonstrate that SVM consistently

DOI: 10.54254/2755-2721/87/20241563

outperforms the mentioned ML approaches, when it comes to both accuracy of classification, measured by Accuracy, and the ability to capture the structural complexity of the data. This becomes evident if we take a look at the performance of logistic regression, which achieved accuracy of 85% and AUC-ROC of 0.88 as well as decision trees model with accuracy of 83% and AUC-ROC of 0.86 [7]. These results showcase the SVM's ability to handle the complexities of financial data effectively.

4. Results and Analysis

4.1. Model Performance

This SVM model has an accuracy of 92%, precision of 89%, recall of 85%, and F1-score of 87% on the test dataset. The value of AUC-ROC is 0.95, which means a high discriminative power and stability of the model. Overall, it shows a model with good prediction performance on defaulted risk of borrowers. A balance setting can be seen among evaluation metrics. The precision is as high as 89%, which means a low rate of false positive [8]. The high value of recall shows the model intended to find most of the defaulters, and it did indeed. For example, in 10,000 people our model classified into defaulter category, 8,500 people are really default the loan. Table 2 visualises each metric's value, what it means, and an example to help understand the model accuracy of default borrower risk.

Table 2. Model Performance Metrics

Metric	Value (%)	Description	Example
Accuracy	92	Proportion of correctly classified instances	92% of all instances were correctly classified
Precision	89	Rate of true positives among the predicted positives	89% of instances predicted as defaults were actual defaults
Recall	85	Rate of true positives among the actual positives	85% of actual defaults were correctly identified
F1-score	87	Harmonic mean of precision and recall	Balanced measure of precision and recall
AUC-ROC	95	Discriminative power of the model	High effectiveness in distinguishing between defaulters and non-defaulters

4.2. Feature Importance

Analyse the feature importance and find out the top three features as borrower income, credit score and debt-to-income ratio as predictors of the default risk. With a closer look, borrower income is the most important feature. As borrower income goes up, the default risk goes down. This could be intuitively reasonable. Borrower credit score reflects a person's creditworthiness directly. Higher the score, lower is the default risk.[9] This feature plays a significant role. Debt-to-income ratio shows how much the borrower is tied up with the financial repayment. It reflects the degree of financial strain. Therefore, a higher debt-to-income ratio, means greater default risk. These results are credit risk management knowledge and empirical evidence in related literature. For example, the default rate is 2 per cent for borrowers with a greater than $50,000 income. Compare to a default rate of 15 per cent with less than $30,000 income.[10].

5. Conclusion

With the help of the large set of borrower attributes and historical loan outcomes, this study has demonstrated that the SVM models can well predict borrower's default risk. The result suggests that SVM models indeed can achieve higher prediction precision than traditional statistical methods and are more resistant to outliers. Current empirical results indicate that financial institutions can definitely benefit from using SVM models for better management of default risk. Insights from this paper suggest that if financial institutions are able to better identify riskier borrowers, they can make better lending decisions (e.g., not to lend money to those who cannot repay), set more realistic interest rates, and allocate more resources to low-risk borrowers. As a result, financial institutions can not only reduce the number of defaults but also maximise profit. More importantly, the relationship between borrowers and financial institutions are strengthened because borrowers do not have to rely as much on private

DOI: 10.54254/2755-2721/87/20241563

intermediaries such as money lenders and have to bear less interest surcharge. Furthermore, when AI models like SVM are used, the credit assessment process can be automated and the loan approval process can be expedited. This allows financial institutions to free up time and reduce payroll expenses related to manual credit evaluation.

References

[1] Roy, Atin, and Subrata Chakraborty. "Support vector machine in structural reliability analysis: A review." Reliability Engineering & System Safety 233 (2023): 109126.

[2] Kurani, Akshit, et al. "A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting." Annals of Data Science 10.1 (2023): 183-208.

[3] Tarzanagh, Davoud Ataee, et al. "Transformers as support vector machines." arXiv preprint arXiv:2308.16898 (2023).

[4] Alhussan, Amel Ali, et al. "Facial Expression Recognition Model Depending on Optimized Support Vector Machine." Computers, Materials & Continua 76.1 (2023).

[5] Durango-Gutiérrez, Maria Patricia, Juan Lara-Rubio, and Andrés Navarro-Galera. "Analysis of default risk in microfinance institutions under the Basel III framework." International Journal of Finance & Economics 28.2 (2023): 1261-1278.

[6] Durango-Gutiérrez, Maria Patricia, Juan Lara-Rubio, and Andrés Navarro-Galera. "Analysis of default risk in microfinance institutions under the Basel III framework." International Journal of Finance & Economics 28.2 (2023): 1261-1278.

[7] Avramidis, Panagiotis, Ioannis Asimakopoulos, and Dimitris Malliaropulos. "Disrupted Lending Relationship and Borrower's Strategic Default." Journal of Financial Services Research 63.1 (2023): 91-116.

[8] Madeira, Carlos. "Adverse selection, loan access and default behavior in the Chilean consumer debt market." Financial Innovation 9.1 (2023): 49.

[9] Bhatt, Tribhuwan Kumar, et al. "Examining the determinants of credit risk management and their relationship with the performance of commercial banks in Nepal." Journal of risk and financial management 16.4 (2023): 235.

[10] Bagale, Sita. "Credit risk management and profitability of commercial banks in Nepal." International Journal of Finance and Commerce 5.1 (2023): 60-67.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241443

Design of a cybersecurity defense system based on big data and artificial intelligence

Minbin Yang

Lingshan Vocational and Technical School, Qinzhou City, Guangxi Province, China

124105694@qq.com

Abstract. With the development of big data and artificial intelligence technologies, hacker attack techniques and capabilities have continuously improved, and the methods of network attacks have diversified. Without cybersecurity, there is no national security. Therefore, cybersecurity has become a focal point of attention. To comprehensively enhance the network security defense capabilities of computer operating systems, aligning with the rapid development trends of big data and artificial intelligence technologies, this study focuses on constructing an efficient, stable, and practical cybersecurity defense system. This system deeply integrates advanced technologies of big data analysis and artificial intelligence, thoroughly analyzing the current status of network information security in computer operating systems and closely aligning with the practical needs of design and production. The aim is to provide highly valuable reference solutions for the field of cybersecurity defense.

Keywords: Big Data, Artificial Intelligence Technology, Computer Networks, Security Defense, System Design

1. Introduction

With technological advancements, network information security issues have become a focal point of public concern. Currently, technologies related to network information security, such as early warning technology, security strategy technology, and continuous network monitoring technology, are relatively lagging. This has resulted in traditional computer network defense techniques and systems being unable to effectively resist such intrusion attacks, with a high rate of missed detections. Considering the service environment of big data and artificial intelligence enterprises, there is an urgent need for updated and improved network information security defense systems. The rise of artificial intelligence technology has provided the most feasible and quickest solutions to computational problems across various industries, especially with the development of core artificial intelligence technologies such as machine learning and deep neural networks [1]. Overall, based on this background, this study is dedicated to designing and constructing a computer network security defense system leveraging the characteristics and advantages of big data and artificial intelligence technologies to assist in solving network information security issues. The content of this paper aims to provide a reference for the design of related cybersecurity defense systems, striving to build a complete and practical defense system to comprehensively enhance the level of computer network information security.

DOI: 10.54254/2755-2721/87/20241443

2. Current Status of Computer Network Security in the Big Data Era

2.1. Increasing Sophistication of Hacker Attack Techniques and Capabilities

Entering the new era, China's intelligent mobile cloud technology has also continued to develop, with emerging technologies such as artificial intelligence beginning to permeate various aspects of our lives. This has led to more people independently accessing and learning about these fields and technologies, gradually mastering more network technologies. Among these groups, there are still individuals with weak legal awareness, some even violating laws for profit. They start using illegal software to attack websites with potential security vulnerabilities, exploiting these system loopholes to obtain personal privacy information of internet users, thereby posing a threat to network security.

2.2. Diversification of Network Attack Methods

In today's world, the mobile internet is developing and being applied at an unprecedented speed, and computer network technology is becoming increasingly complex. The number of intelligent mobile terminals on network platforms is also continuously expanding. These technologies are not only applied to common devices and equipment such as smartphones, tablets, and laptops but also to many large and complex electronic digital products. While this has made people's lives more convenient and work easier, it has also laid a solid foundation for network security threats, increasing the difficulty of cybersecurity defense.

3. Requirements for the Design of a Cybersecurity Defense System Based on Big Data and Artificial Intelligence

In the current design of network information security defense systems for computer operating systems, it is crucial to approach from a practical perspective, considering ways to effectively enhance the security defense level. During the design phase, comprehensive planning is essential, understanding the system architecture and actual layout [2]. Specifically, the following measures need to be taken to address these requirements:

Timely Response and Accurate Judgment: When computer network technology faces external illegal intrusions, the constructed security defense system must quickly respond, transmitting information to the intrusion point. This enables clear identification of the type and purpose of the intrusion. However, achieving comprehensive effectiveness in reducing recognition accuracy and false alarm rates is challenging.

Real-time Automatic Response and Digitalization, Intelligence: The system must respond promptly to different security events after the program runs and handle events through automatic shutdown analysis. Digitalization and intelligence significantly improve the efficiency of security handling and make system operation more automated.

Intrusion Tracking and Non-response to Threats: The system should automatically shut down the tracking of large-scale intrusion behaviors in the computer operating system and respond to factors that may threaten the system's safety during the operation of the computer host programs.

4. Design and Implementation of a Cybersecurity Defense System Based on Big Data and Artificial Intelligence

4.1. System Intrusion Detection and Alarm Module

The primary task in designing and producing information security defenses and systems for computer operating systems is to leverage the development of big data and artificial intelligence technologies. The intrusion detection system alarm module has robust functionalities, quickly identifying relevant information and database data after the computer operating system is attacked. Integrating the development of artificial intelligence technology with high-performance detection sensors works synchronously to further enhance the speed of the custom module's red alert. Figure 1 illustrates the specific design of the intrusion detection system alarm module.

DOI: 10.54254/2755-2721/87/20241443

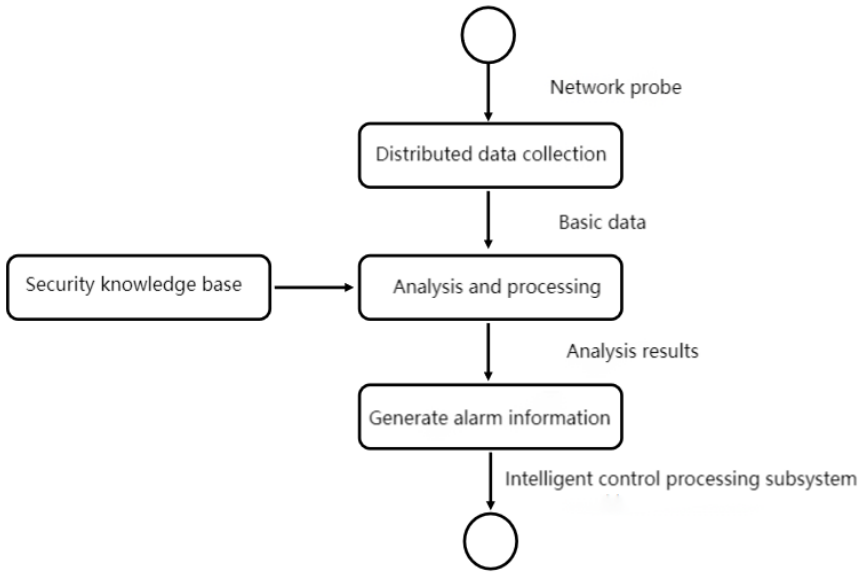

Figure 1. Design and Implementation of the System Intrusion Detection and Alarm Module

While ensuring that alarm detection and calls are conducted as scheduled, the system must also support the efficient and stable operation of eavesdroppers in different network segments. In the detailed data analysis process of detection data, the following high-performance technologies are utilized, such as analyzing user behavior and studying precise protocol content. Additionally, external resources can be collected as much as possible for better software detection. By applying specific event analysis technologies, the time for transmitting information to the intelligent control system's custom module for processing is minimized, achieving unique event analysis security defense. This further enhances the accuracy, stability, and reliability of software detection data.

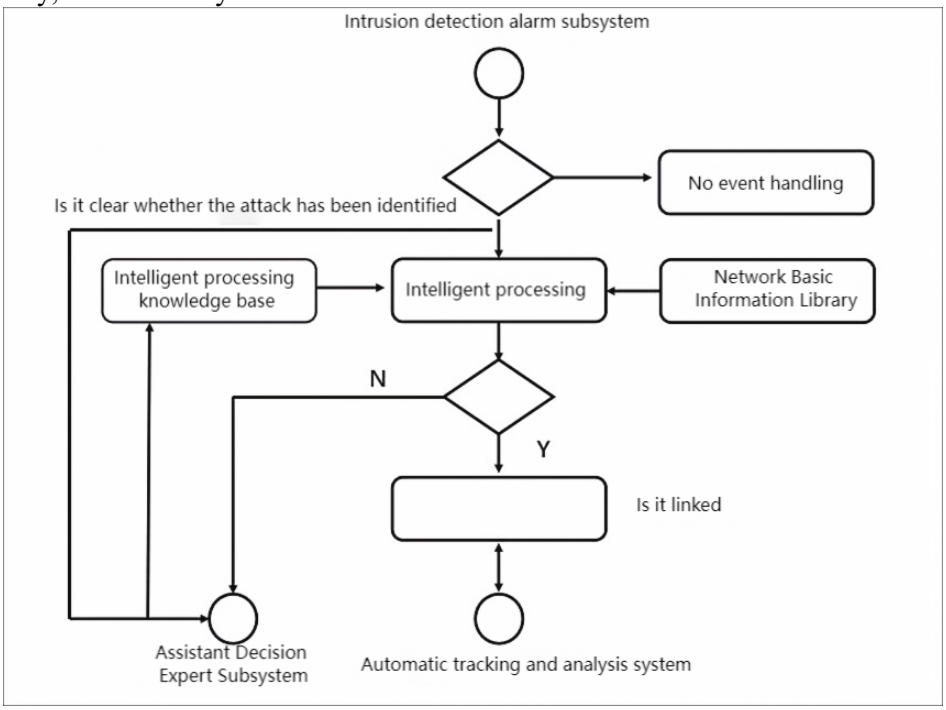

Figure 2. Design and Implementation of the System Intelligent Processing Module

DOI: 10.54254/2755-2721/87/20241443

4.2. System Intelligent Processing Module

The intelligent processing module of the cybersecurity defense system is used to receive security defense-related information collected by the custom alarm module of the intrusion detection system. Its main task is to determine the conditions under which a fixed IP address is under attack and immediately disconnect the network connection associated with that network port to prevent further large-scale intrusions from damaging the computer operating system.

The basic working principle of the intelligent processing module in the cybersecurity defense system for specific situations is shown in Figure 2.

1. Emergency Action: This type of defensive action requires the computer network technology to respond to intrusion situations as quickly as possible to stop the intrusion promptly. The goal of "urgent action" is to take the most direct corrective measures as quickly as possible.

2. Timely Action: When an intrusion occurs, under a timely action control mechanism, the system may not respond to how to handle the intrusion, which could extend for days. The goal of avoiding wasted time must be accomplished through a custom intelligent control system.

3. Local Long-term Action: This type of action is relatively less severe compared to previous situations but aims to be as detailed as possible, allowing security defense personnel to analyze and organize the information.

4. Global Long-term Action: Compared to previous forms, global long-term actions involve the entire computer operating system. In long-term global actions, stricter requirements and specific criteria are introduced for the entire perimeter of the network system. In the system, the response system components and the two custom modules for authorized expert decision support and automatic shutdown tracking analysis maintain a close and normal connection regardless of whether the transmission of relevant banking data is supported.

The specialized custom modules that fully provide decision support are usually connected to the analysis system components and, for various reasons, to the data collection system components for system intrusion detection system alarm calls. This ensures efficient and stable transmission of relevant information and data between system components. Moreover, during theoretical and practical processes, system technology comparisons can support the most critical aspects of system defense, facilitating data sharing and recovery work between them, thus reducing losses caused by the operation of the network system.

4.3. System Auxiliary Decision-Making Expert Module

When designing and producing information security defenses and systems in computer operating systems, the development of big data and artificial intelligence technologies should be combined to design an expert-defined system with complete modules that provide reliable decision support. The primary task of this custom module is to automatically generate suggestions and optimal plans in case of specific intrusions and alarms to assist system security-related managers, forming more optimized results for system security. For decision-makers, it provides excellent support and solutions [3]. When designing and producing the expert-defined module for security and decision support in computer operating systems, reference can be made to the specific content in Figure 3. The knowledge graph's security knowledge is an essential part of the custom professional module, fully providing decision support authority and storing all specific safety knowledge content in the retrieval system. The designed defense and system can quickly take all defense response methods when facing intrusions [4]. Additionally, the custom module that fully provides decision support has a robust automatic shutdown learning function, providing more intelligent decisions for the subsequent security defense of the computer operating system during the continuous learning process.

DOI: 10.54254/2755-2721/87/20241443

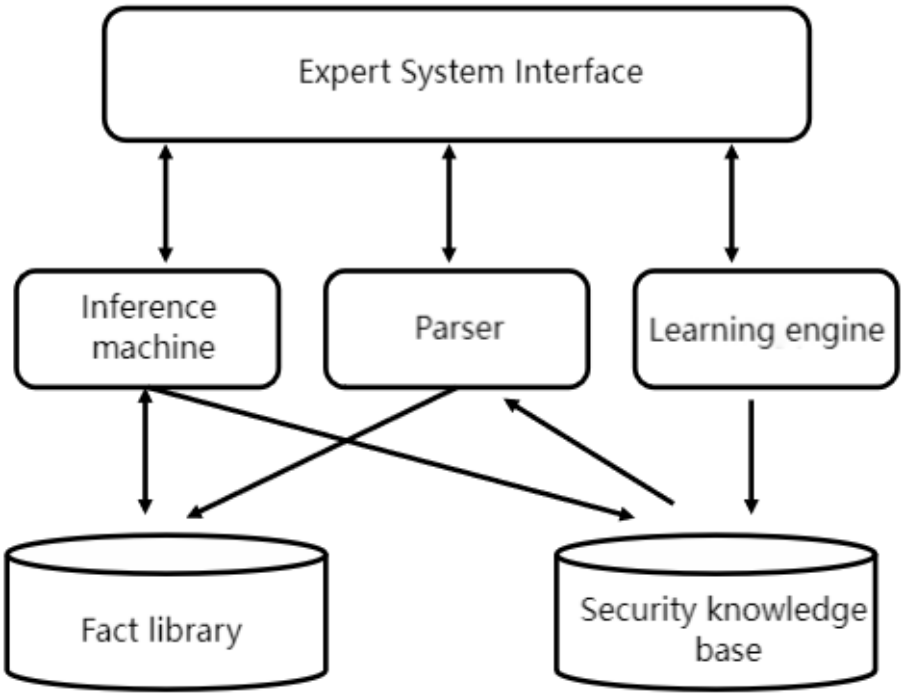

Figure 3. Design and Implementation of the System Auxiliary Decision-Making Expert Module

4.4. System Automatic Tracking Analysis Module

Using automated, in-depth analysis to address crises, the system will automatically shut down and initiate the tracking analysis custom module, taking proactive remedial measures to counteract intrusions when the information source or the entire specific information flow is attacked [5]. This custom module also helps to improve and optimize subsequent cybersecurity defense plans for computer operating systems.

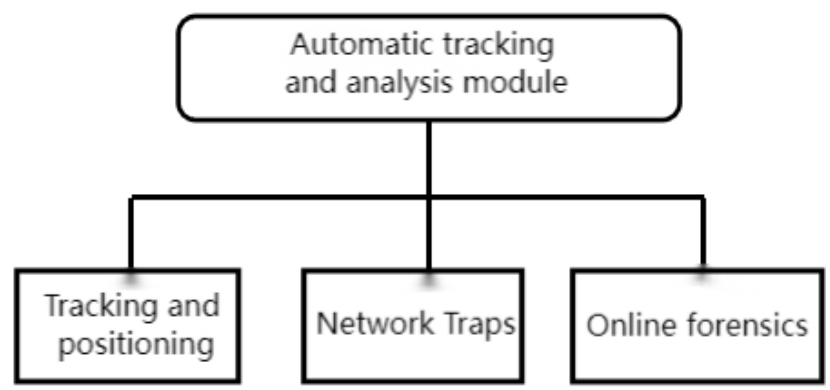

Figure 4. Design and Implementation of the System Automatic Tracking Analysis Module

Based on the basic design and production structure shown in Figure 4, it can be concluded that these sub-custom modules operate independently but within the computer operating system's security protection system. They must be connected to the custom module of the intelligent control system to collaborate effectively.

DOI: 10.54254/2755-2721/87/20241443

5. Conclusion

In summary, to further improve the quality of network information security work in our country's computer operating systems in the new era, we need to invest effort in developing big data and artificial intelligence technologies. The most critical task in constructing network information security defenses and systems is to build custom modules for intrusion detection system alarms. This allows local network-connected computer users to promptly understand and detect intrusion points and specific intrusion details, thus enhancing the security system and improving its efficiency and quality.

Author Biography: Yang Minbin, born in December 1986, female, native of Lingshan, Guangxi; Education: Bachelor's degree; Workplace: Qinzhou Lingshan Vocational and Technical School; Position: Vice Principal; Title: Senior Lecturer; Research direction: Computer technology; Email: 305615514@qq.com.

References

[1] Zhang Xiaoyan. Application of Artificial Intelligence Technology in Cybersecurity Defense [J]. Information System Engineering, 2021(07): 58-60.

[2] Zhang Rong. Research on a Multi-level Cybersecurity Defense Model Based on Artificial Intelligence [J]. Information & Computer (Theory Edition), 2021(13): 180-182.

[3] Liao Yuxiang. Application of Artificial Intelligence Technology in Cybersecurity Defense [J]. Information Technology and Informatization, 2021(06): 182-184.

[4] Wang Yang. Analysis of Information Security Risks and Preventive Measures in the Context of Big Data [J]. Cybersecurity Technology and Applications, 2020(11): 9-11.

[5] Li Fei. Optimization Strategies for Computer Network Security Technology in the Big Data Environment [J]. Computer and Information Technology, 2020, 28(5): 66-68.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241519

Research on the performance of hybrid vision models based on ViT

Bowen Chai

Shanghai University of International Business and Economics, School of International Business, Shanghai, 201620, China

balwyn134821589341@163.com

Abstract. ViT is a model proposed by the Google team in 2020 to apply a transformer in image classification, although it is not the first paper to apply a transformer in visual tasks, because of its model is "simple" and effective, and scalable, it has become a milestone work in the application of transformer in CV field, and has also triggered the subsequent related research. The core conclusion of the original ViT paper is that when there is enough data for pre-training, ViT outperforms CNN, breaks through the limitation of the transformer's lack of inductive bias, and can achieve better migration results in downstream tasks. However, when the training dataset is not large enough, ViT usually performs worse than ResNets of the same size, because Transformer lacks inductive bias, a kind of a priori knowledge, assumptions made in advance, compared with CNN. Through its innovative architecture and powerful performance, the visual representation transform (ViT) model continues to advance the field of computer vision, while facing some challenges and room for improvement. With the deepening of research and the continuous development of technology, ViT is expected to play a greater role in more practical applications. The article aims to explore the advantages and applicability of the ViT model and tries to construct a hybrid visual model to improve its generalization ability for different types of datasets, demonstrating the hybrid model's significance in improving the performance of the ViT model.

Keywords: Hybrid Vision Models, Vision Transformer, CIFAR10, CNN+ViT, Performance

1. Introduction

The development of ViT can be traced back to 2017 when the Attention Is All You Need paper proposed the transformer structure for realizing machine translation tasks.[1] Later, the transformer structure was widely used in speech recognition, natural language processing and other fields, and it was also explored and attempted in the image field. In 2018, some scholars proposed to use a transformer instead of CNN, which was called Image Transformer, but the effect was not as good as CNN. [2]Until 2020, the ViT model proposed by the Google team[3] used the transformer structure to realize the image classification task, and achieved comparable performance with the CNN model, which is called "Vision Transformer".[4]

According to He et al, ResNet50 is a convolutional neural network model with 50 convolutional layers. It was proposed by researchers at Microsoft and won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2016. The model adopts the idea of residual learning to avoid the

DOI: 10.54254/2755-2721/87/20241519

gradient vanishing problem of deeper networks while ensuring the expressive power of deeper networks.[5]The principle of ResNet50 is mainly to fuse the feature maps of the previous layers with those of the later layers through direct connections across channels, so that the whole model can be deeper and maintain a certain gradient flow, avoiding the deep network's gradient vanishing problem. Meanwhile, techniques such as batch normalization, pre-activation and residual block are also used to further improve the expressive ability and training speed of the model.

The CIFAR-10 dataset is open-access and is often used to train and evaluate computer vision algorithms. Since the images in the dataset are small, the algorithms can be trained and tested quickly, and it is suitable as an entry-level dataset for training and testing some basic computer vision algorithms, such as object recognition, classification, localization, tracking, and other tasks. Another important significance of the CIFAR-10 dataset is that it has also become a standard benchmark dataset for the evaluation of some deep learning algorithms, such as ResNet50, Inception, etc. Part of CIFAR-10 is shown in Fig. 1:

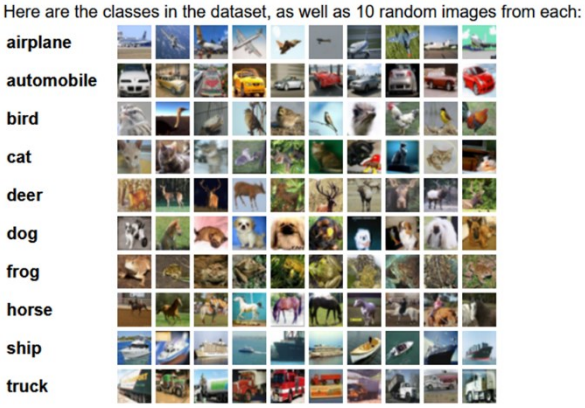

Figure 1. Selected images in CIFAR-10

This paper explores the performance of ViT model in small-scale image classification tasks by comparing the performance of the traditional ViT model and ResNet50 on the CIFAR10 dataset. Since the performance of ViT is slightly inferior in dealing with small-scale datasets, this paper will improve the traditional ViT model by introducing a variable convolutional layer to construct a CNN-ViT model, and further compare the performance of the three models on the CIFAR10 dataset.

Compared with the existing literature, the possible marginal innovations of this paper are: (a) Introducing variable convolutional layer to construct CNN-ViT model, in this paper, CNN-ViT model is constructed by introducing variable convolutional layer, which mixes convolutional neural network and ViT model, to give full play to the advantages of each of them. Compared with the traditional ViT model, the CNN-ViT model has better feature extraction capability and higher classification accuracy and shows better performance when dealing with a variety of datasets such as the CIFAR10 dataset. (b) Constructing hybrid visual models to improve generalization ability, in this paper, the generalization ability of the ViT model on small-scale datasets is improved by constructing hybrid visual models. The hybrid model fully exploits the inter-correlation between data to enhance the generalization ability of the model in dealing with small-scale datasets where ViT is not originally dominant. The results show that the hybrid visual model plays an important role in improving the generalization ability of the ViT model.

2. Research Methodology

2.1. Selection of Dataset

In this study, many factors are considered and finally, CIFAR-10 was decided to be chosen as the training dataset.

DOI: 10.54254/2755-2721/87/20241519

According to Chen, Deng, & Du, large-scale datasets commonly used for visual model training are ImageNet, JFT-300M, Google Landmarks v2, iNaturalist 2018, etc.[6] They contain more than one million image data. However, limited by hardware devices and network conditions, finally the relatively small-scale CIFAR-10 was chosen as the training dataset. Although the ViT model is not quite able to demonstrate its advantages on such small-scale datasets, this paper subsequently tries to improve the traditional ViT model based on this and explores whether it can enhance its performance on small-scale datasets.

2.2. Training the Vit Model

The data comes from the public dataset on the web. The code in this paper is borrowed from Chatgpt and GitHub-related open-source code.[7]

The first step is to import the PyTorch library and obtain the ViT model code.

Step 2: Define the image classifier, training function and test function.

Step 3: Load the dataset

Step 4: define hyperparameters and optimizer

Step 5: start training. The number of iterations of the training process is controlled by the epochs variable and traverses the dataset loader in each iteration for each batch of data. In each training iteration, the code also calculates the average loss value, accuracy, and total number of samples in the training set, and after using optimizer.zero_grad() to perform a zero operation on the model gradient, it uses the backwards () gradient backpropagation function,[8] which computes the gradient value corresponding to each sample point, and then uses scaler.step() to the parameter in the optimizer to update it. Finally, the code also counts the information such as the amount of correctness between the output results and the actual labels and the total amount of correctness in each batch of data to calculate the accuracy of the training, and stores the information such as the total loss degree and the accuracy of that training, respectively, in a list to visualize the training process after the training is completed.

2.3. Training the Resnet50 Model

In the first step, TensorFlow [9] and Keras [10] libraries are used for the construction of the convolutional neural network model, which performs the classification task for ten different categories of the CIFAR-10 image dataset.

In the second step, the CIFAR-10 dataset is imported.

In the third step, the labels of the dataset are uniquely thermally encoded, converted into a vector consisting of 10 binary bits, and assigned to the y_train and y_test variables, respectively. For the image data in the x_train and x_test variables, normalization is performed before model training is performed by converting the pixel values from integers from 0 to 255 to floating-point numbers between 0 and 1 and preprocessing the data into a format that meets the requirements of model training.

In the fourth step, the construction of the ResNet50 model was started, using the ResidualBlock function to implement the residual block.

Finally, the Adam optimizer is set up with a loss function of categorical_crossentropy and an evaluation metric of accuracyaccuracy; the best model is saved periodically for the validation set during model training through the ModelCheckpoint and EarlyStopping callback functions, and training is stopped early to prevent overfitting Use the fit function to train the model, pass in the training set and test set, specify the batch size as 128 and the number of training periods as 20 and call the callback function during the training process to complete the training of ResNet50 for CIFAR-10 and output the training results.

2.4. CNN+ViT Model Construction and Training

The idea of constructing a hybrid CNN+ViT model to improve the performance of ViT comes from an article on ViT parsing posted on the web by a scholar from Zhejiang University.[11] According to Zhang, J., the core conclusion of the original ViT paper is that when there is enough data for pre-training, ViT outperforms CNN, breaks through the limitation of the transformer's lack of inductive bias, and can

DOI: 10.54254/2755-2721/87/20241519

obtain better migration results in downstream tasks. However, when the training dataset is not large enough, ViT usually performs worse than ResNet of the same size, because Transformer lacks inductive bias, a kind of a priori knowledge, and assumptions made in advance, compared with CNN, which makes CNN have a lot of a priori information and need relatively less data to learn a better model.

Then, since CNN has the property of inductive bias, and the Transformer has strong global inductive modelling ability, perhaps a hybrid model using CNN+Transformer can get better results. So, this paper constructs a CNN+ViT hybrid model and continues to use it to train CIFAR-10, and finally compares the training performance of these three models.

Again, the PyTorch open-source library Timm was imported first, so that the ViT model could be imported later.

Next, a three-layer convolutional neural network (CNN) was imported to work with the original ViT model.

The ViT encoder uses the Transformer to process the features generated by the CNN encoder, which splits the image into blocks of a pre-set "patch_size", in this case, 64, and embeds each block into a low-dimensional vector representation. These vectors are then processed under the multi-head self-attention mechanism and sent to the fully connected network for final classification.

After this again the CIFAR-10 dataset is imported and preprocessed; and the Adam optimizer and loss function are defined and training is started to record the results.

3. Experimental Results Analysis

This project used Google Colab to complete the training of the models, ran in Python 3 Google Compute Engine backend (GPU) mode, and used Tableau to visualize the training results in data. All models were trained with Epoch preset to 20.

Among them, the ViT model was used to train CIFAR-10 20 times, ResNet50 3 times, and CNN+ViT 10 times; the more complete data of them was recorded, as shown below:

3.1. Training Results of the Vit Model

As shown in Table 1 and Figure 3, the training results of the ViT model are presented as follows, and the training duration is 7 hours.

Table 1. ViT training CIFAR-10

Epoch	Training Set Loss Degree	Training Set Accuracy	Test Set Loss Degree	Test Set Accuracy
1	1.6349	27.53%	0.0136	35.71%
2	1.5205	38.20%	0.0128	41.08%
3	1.2647	45.46%	1.4291	46.88%
4	1.3445	49.24%	0.0109	50.12%
5	1.4079	51.70%	0.0107	51.36%
6	1.2204	52.97%	0.0101	54.35%
7	1.2554	54.05%	0.0099	54.11%
8	1.2286	55.30%	0.0098	54.59%
9	1.2731	56.35%	0.0097	56.13%
10	1.2527	56.85%	0.0091	57.03%
11	1.2284	57.61%	0.009	58.54%
12	1.3317	58.12%	0.0088	59.17%
13	1.0816	58.97%	0.009	58.41%
14	1.192	59.26%	0.0085	60.56%
15	1.2002	60.30%	0.0085	60.97%
16	1.2811	61.17%		

DOI: 10.54254/2755-2721/87/20241519

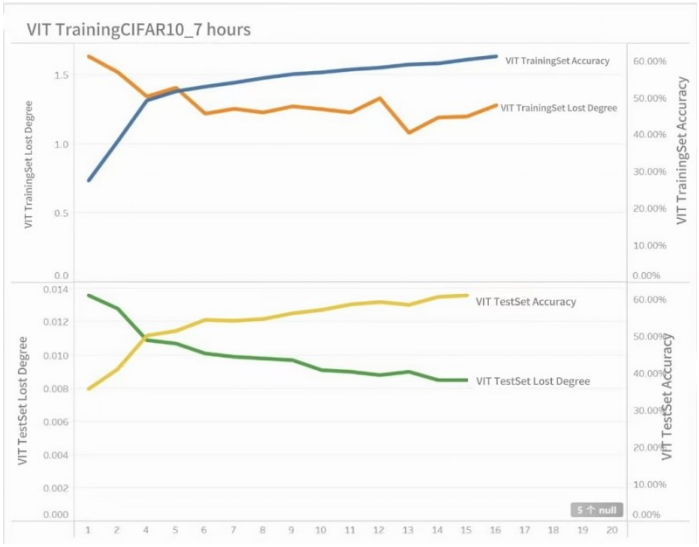

Figure 3. Visualization of ViT model training results

3.2. Training Results of ResNet50 Model

As shown in Table 2 and Figure 4, the training results of the ResNet50 model are presented as follows, with a training duration of 6 hours and 40 minutes.

Table 2. ResNet50 training CIFAR-10

Epoch	Training Set Loss Degree	Training Set Accuracy	Test Set Loss Degree	Test Set Accuracy
1	1.8700	43.07%	2.1633	26.16%
2	1.3869	57.82%	5.1761	22.96%
3	1.2296	63.59%	2.025	32.55%
4	1.0266	63.59%	1.2364	60.98%
5	0.8755	72.25%	0.862	69.77%
6	0.7357	75.79%	0.9031	70.63%
7	0.6806	77.34%	1.0124	66.27%
8	0.5295	81.63%	0.7192	76.26%
9	0.4534	84.19%	0.8596	73.39%
10	0.3787	86.76%	0.715	77.48%
11	0.3157	88.83%	1.3259	67.83%
12	0.3037	89.51%	0.9077	73.80%
13	0.2145	92.38%	1.0554	73.75%
14	0.1727	93.87%	0.9612	75.96%
15	0.153	94.55%	0.8413	79.21%
16	0.1267	95.54%	1.0632	77.02%
17	0.1129	96.00%	1.2959	74.89%
18	0.1037	96.39%	1.0033	77.04%
19	0.0957	96.70%	0.9779	78.35%

DOI: 10.54254/2755-2721/87/20241519

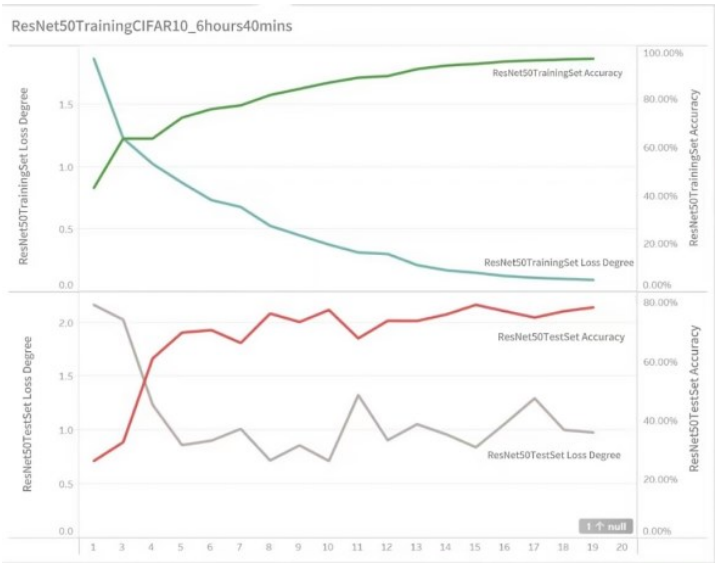

Figure 4. Visualization of ResNet50 training results

3.3. Training Results of CNN+Vit Model

As shown in Table 3 and Figure 5, the training results of the ResNet50 model are presented as follows, and the training time is 4 hours and 17 minutes.

Table 3. Training results of CNN+VIT

Epoch	Hybrid Model Loss Degree	Hybrid model correctness
1	1.9532	40.35%
2	1.4583	50.60%
3	1.2619	56.85%
4	1.1469	60.48%
5	1.0673	62.56%
6	0.9989	63.56%
7	0.9459	67.12%
8	0.9056	68.60%
9	0.8577	69.10%
10	0.799	70.28%

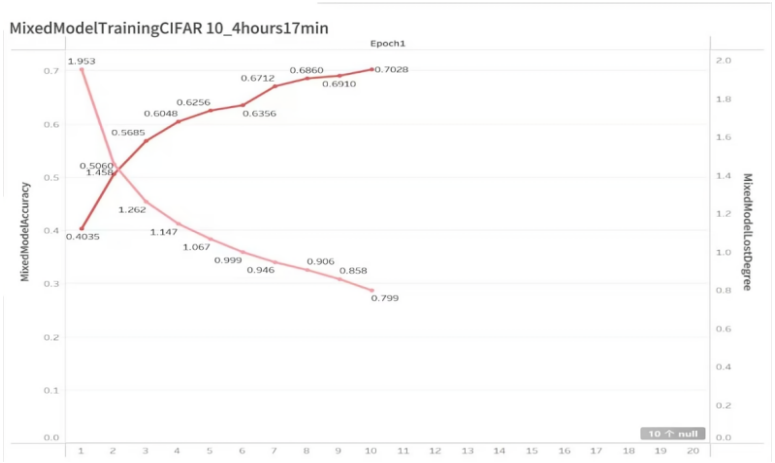

Figure 5. Visualization of CNN+ViT training results

DOI: 10.54254/2755-2721/87/20241519

3.4. Comparison of the Three Models

As shown in Table 4, the following results can be obtained from the comparison:

Table 4. Training results for each model

Model name	Training hours	Total traversals	Maximum accuracy (test set)
ViT	7hrs	16	60.97%
ResNet50	6hrs 40mins	19	78.35%
CNN+ViT	4hrs 17mins	10	70.28%

As we can see, CNN+ViT has exceeded the accuracy of the normal ViT model despite having the lowest number of traversals, ResNet50 has the highest number of training traversals and also has the highest correctness rate, it is believed that if CNN+ViT can achieve the same number of training layers, its performance can be close to that of ResNet50 or even exceed it.

As shown in Touvron et al., ResNet50 outperforms ViT on small datasets such as CIFAR-10, while ViT performs better on larger datasets such as ImageNet-1k with more than a million images.[12]

Figure 6 presents the correctness of the three models in the form of curves.

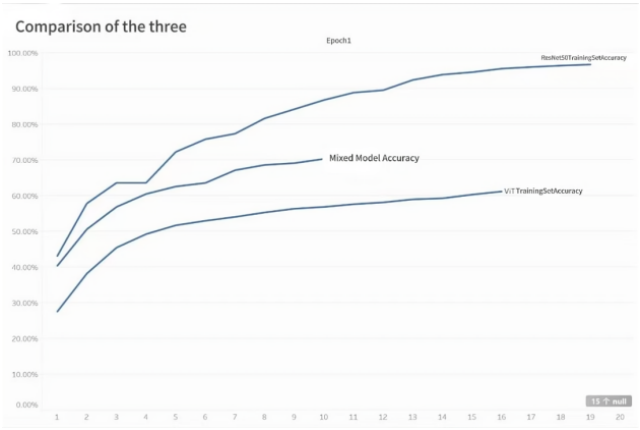

Figure 6. Comparison of the three models' accuracy

4. Conclusion

This paper explored the performance metrics such as the number of training traversals and accuracy under different neural architectures. It is found that the CNN+ViT model has a lower number of traversals (only ten), but its accuracy has surpassed that of the regular ViT model.Meanwhile, the ResNet50 model has the highest number of training traversals, but also exhibits the highest correctness rate. The CNN+ViT model is believed that if it can achieve the same number of training layers as ResNet50, its performance will keep approaching or even surpass the ResNet50 model. Our results are in line with Touvron et al. that the ResNet50 model performs well on small datasets such as CIFAR-10, while the ViT model performs better on much larger datasets (e.g., ImageNet-1k, which has more than one million image data).

Not coincidentally, according to Zheng et al., in a small-sample image classification task, they constructed a hybrid model of a Convolutional Neural Network and Transformer called LCPN and explored the synergy between local and global features.[13] The article proposes a Local Composition Module (LCM), which has a structure very similar to a Transformer but does not use the attention mechanism to process local information. Meanwhile, the model combines LCM and CNN, which makes LCPN can classify small samples more effectively. Through extensive experimental evaluations on several small-sample image classification benchmark datasets, the article demonstrates that LCPN has excellent performance in image classification tasks with small samples and has an advantage in innovatively capturing information about local compositional patterns.

DOI: 10.54254/2755-2721/87/20241519

And recently there have been many new models derived based on the ViT model, such as ViT-pyramid.[14] the model adopts a pyramidal multi-scale feature representation, i.e., the input image is scaled to different scales, and then features are extracted at each scale, which are finally fused to classify the image. the main idea of ViT-pyramid is to replace the global field of view with a finer field of view, to improve the classification accuracy and efficiency of the model. thereby improving the classification accuracy and efficiency of the model. Its core structure is a set of Transformer Block-based feature extraction layers, each of which contains a self-attention mechanism for learning the relationship between the current position and other positions. These layers are characterized by the ability to dynamically adjust the size of the region of the self-attention mechanism to achieve feature extraction and combination at different scales. ViT-pyramid has achieved good performance on several datasets, especially on large-scale datasets, such as ImageNet-21K.

U-ViT, proposed in CVPR2023,[15] combines Vision Transformer, a vision model, with U-Net, and applies it in Diffusion Model (Diffusion Model) to replace the original CNN, and applies the long skip structure of U-Net in Transformer as well to realize image generation using the Transformer for the task of image generation.

References

[1] Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez & Polosukhin (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[2] Siva, Wong & Gong (2018). Improved techniques for training GANs. In Proceedings of the 35th International Conference on Machine Learning (ICML), Stockholm, Sweden (pp. 5070-5079).

[3] Alexey Dosovitskiy. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale [EB/OL]. [2023/5/27]. https://arxiv.org/abs/2010.11929.

[4] Dosovitskiy, Beyer, Kolesnikov, Weissenborn, Zhai, Unterthiner, & Houlsby, (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR).

[5] He, Zhang, Ren & Sun, (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[6] Chen, Deng & Du (2021). Recent Advances on Vision Transformers. arXiv preprint arXiv:2106.13129.

[7] Hugging Face. (n.d.). pytorch-image-models/vision_transformer.py. GitHub. Retrieved June 3, 2023, from https://github.com/huggingface/pytorch-image-models/blob/main/timm/models/vision_transformer.py

[8] PyTorch. (n.d.). Optim: Implementations of Gradient Descent Algorithms. PyTorch Documentation. Retrieved June 3, 2023, from https://pytorch.org/docs/stable/optim.html.

[9] TensorFlow. (2021). TensorFlow documentation. Retrieved October 25, 2021 from https://www.tensorflow.org/

[10] Keras. (2021). Keras documentation. Retrieved October 25, 2021 from https://keras.io/

[11] Zhang, J. (2021). PyTorch Lightning 1.2 - New Modules and Features. Retrieved September 7, 2021, from https://zhuanlan.zhihu.com/p/445122996?utm_id=0.

[12] Touvron, Caron, Alayrac & Misra (2021). From ResNets to Pre-Training: Revisiting ImageNet Pre-Training. arXiv preprint arXiv:2105.14333.

[13] Zheng, Wei, Yang, Zhang & Huang (2021). Exploiting Local Compositional Patterns for Few-Shot Image Recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5908-5917.

[14] Tolstikhin, Houlsby, Kolesnikov, Beyer, Zhai, Unterthiner & Raffel, C. (2021). Multi-scale Vision Transformers: An Evolution towards Competitive Computer Vision Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14352-14361).

[15] Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., & Zhu, J. (2023). All are Worth Words: A ViT Backbone for Diffusion Models. arXiv preprint arXiv:2209.12152v4.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241545

The investigation of traditional models and machine learning models in dynamic facial expression recognition

Xiyu Wu

The Department of Mathematics and Statistics, South-Central Minzu University, 430074, China

202221101052@mail.scuec.edu.cn

Abstract. In everyday life, dynamic facial expressions are merely continuous human responses to external events. However, in human-computer interaction, rapidly recognizing changes in facial expressions from video streams is a relatively complex process. This complexity renders Dynamic Facial Expression Recognition (DFER) a critical research task in the domains of computer vision and image processing. This paper analyses the correlations and contrasts between static and dynamic facial expression research, highlighting key issues in the study of dynamic facial expressions, such as dynamic feature extraction and frame extraction. After that, it enumerates significant algorithms in both traditional models and deep learning models, providing an analysis of the advantages and disadvantages of these two major approaches. At the same time, it investigates the reasons behind the transition of research models for DFER from traditional methods to deep learning approaches. The paper focuses on two notable models from each approach: Histogram of Oriented Gradient (HOG) for processing raw images, Support Vector Machine (SVM) for data classification in traditional models. Convolutional Neural Network (CNN) for spatial feature extraction and Long Short-Term Memory (LSTM) for temporal feature extraction in deep learning models. These models are discussed in detail concerning their strengths and weaknesses, operational processes, and performance outcomes. In the concluding section, the author summarizes the main factors influencing research in this field and the current challenges encountered. By focusing on future research directions, the paper also presents a review of recent methodologies and offers insightful research directions for further investigation.

Keywords: Traditional model, machine learning, dynamic facial expression recognition.

1. Introduction

Facial expression is one of the important forms of non-verbal communication in interpersonal communication. It is of great significance to establish good interpersonal relationships and promote effective communication. Since DFER has practical importance in public safety, human-robot interaction, psychological health monitoring and other fields, it has recently received increased attention.

In the 20th century, Ekman and Friesen identified six primary emotions that were disgust, anger, fear, sadness, happiness, and surprise. In later studies, contempt was included in the list of the primary emotions. According to the feature representations, Facial Expression Recognition systems can be divided into two main classes: dynamic sequence FER and static image FER. In the study of dynamic

DOI: 10.54254/2755-2721/87/20241545

fields [1], they considered the time static characteristics of dynamic texture and temporal texture, and their local information and spatial position.

There are some algorithms in traditional models used for FER such as geometric features, template matching, Eigenfaces features algorithms, Hidden Marcov method, singular value decomposition method. While recognition of geometric features is sensitive to the positions of feature points, template matching is sensitive to head posture and scale changes. Additionally, the Eigenfaces method is sensitive to changes in lighting and micro-expressions, while singular value decomposition is sensitive to noise and outliers in the data. All of them will affect the accuracy of features extraction. Meanwhile, these algorithms cost a lot when processing high -dimensional data, which is short of the performance requirements in actual use.

With the new forms of information technology, in order to help computer systems recognize and solve different problems from the surrounding environment, scholars gradually devote themselves to the study of Machine Learning [2] and Deep Learning [3]. The effectiveness of machine learning depends on the integrity of input data, and the feature extraction of facial expressions will be affected by lighting, posture, and obstruction etc. They might lead to prejudice feature selection that may lead to incorrect discrimination between classes. Therefore, using traditional methods including machine learning usually requires manual features extraction, which greatly reduces efficiency. For example, Local Binary Patterns (LBP), LBP on three orthogonal planes (LBP-TOP), Histogram of Oriented Gradient (HOG), Scale-invariant Feature Transform (SIFT), Support Vector Machines (SVM). Relatively, features extraction is automatically implemented in Deep Learning. Convolutional Neural Network (CNN) is an extremely popular approach, which can automatically detect the most distinctive features without any manual supervision. It can also reduce the number of training network parameters to a certain extent, help the network enhance the generalization and avoid overfitting. Artificial intelligence uses this similar layered architecture to simulate the process of the core sensory region of the human brain.

Benefit from the rapid development of DFER, the objective of this paper is to conduct an exhaustive review of the research on DFER through both traditional and machine learning models. At the same time, it introduces the latest research progress. Finally, the challenges that must be addressed to make DFER research applicable to real-world situations are discussed.

2. Method

2.1. Traditional model-based DFER

2.1.1. Histogram of oriented gradient (HOG)-based model

Local objects are characterised by their form and appearance through the distribution of edge directions or local intensity gradients using HOG, a shape descriptor. The main purpose of this program is to detect objects, but it can also be used to intuitively model the shape of facial muscles through edge analysis. Moreover, the input data will not be a factor affecting the parameter configuration of HOG. However, if it is desired to achieve higher detection performance, it is required to adjust the HOG parameters to a fine scale inverse, more direction boxes, and medium-sized, strongly normalized, overlapping descriptor blocks. Existing study has shown that the highest FER performance can be achieved if the parameters are configured to a unit size of 7 pixels and 7 direction boxes [4]. HOG can be subdivided into rectangular HOG (R-HOG) and circular HOG (C-HOG) according to the geometric shape of the descriptor block. In DFER, the frame image is pre-processed using the HOG method, which typically involves converting the image to grayscale. The Sobel operator is then applied to determine the horizontal and vertical gradients. Subsequently, the image is split up into a number of descriptor blocks. In every block, the histogram of gradient directions is computed and the frequency of each gradient direction is documented. To generate the feature vector, the normalized gradient histograms are concatenated in the end.

DOI: 10.54254/2755-2721/87/20241545

2.1.2. Support vector machines (SVM)-based model

The capability to be used for classification and regression prediction can be obtained from SVM, a generalized linear classifier. It is suitable for pattern classification and regression-based applications. Due to its strong statistical basis and effectiveness, it is capable of using linear function hypothesis space in high-spatial feature area. By building a hyperplane which is associated with decision planes in higher dimensions, SVM performs classification. This hyper-plane refers as decision planes, which can make a distinguish between two different groups of data. Data in higher-dimensional spaces is categorized by constructing a hyperplane using a suitable non-linear mapping. The investigation of SVM involves examining the support vectors that determine the decision boundary and yield a significant marginal separation between the classes. SVM distinguishes between classes by recognizing different expression types with the maximum marginal distance [5]. In the study of dynamic facial expressions, the feature vector extracted by HOG can be input into the SVM library, and then the SVM became a useful tool to classify the observed facial emotions. SVM has been enhanced in different ways in recent decades, including Lagrangian SVM, twin SVM, Least Square SVM, Quantum SVM and many others improvements.

2.2. Deep Learning model-based DFER

2.2.1. Convolutional Neural Network (CNN)-based model

Using CNN, the characteristic of countenance can be extracted through a feedforward neural network. The biggest difference from conventional feature extraction methods is that it does not require manual feature extraction and can respond to various features automatically. Its overall architecture includes input layer, folding layer, pooling layer and completely connected layer. In the study of DFER, the biggest difficulty is how to extract effective facial features in the video. Thanks to the convolution layer of CNN, it can continuously abstract the original frame image and extract effective features layer by layer. The multiple convolution cores in the convolution layer can ensure that CNN extracts multiple feature descriptions of facial expressions in each frame during the learning process. Recently, a study proposed a method to combine CNN and HOG to extract more comprehensive dynamic facial expression features [6]. If DFER is performed, ordinary CNN can only obtain the spatial relationship of the input data but not the temporal relationship. To overcome this limitation, the concept of 3D convolutional neural network (3DCNN) was proposed. Large deep CNN can use pure supervised learning. It is vital to note that the depth of CNN is very important for the realization of expression recognition. If a single convolutional layer is removed, its network performance will decrease. Especially when used on video sequences, this urgently requires very large and deep convolutional networks [7].

2.2.2. Long Short-Term Memory (LSTM)-based model

LSTM is an impactful tool for sequentially encoding spatiotemporal features. It was created to alleviate the gradient vanishing or exploding problem encountered by traditional recurrent neural networks when dealing with long-term dependency problems. It replaces the hidden layer of the traditional RNN with a composite unit containing input nodes, input gates, internal states, forget gates, and output gates [8]. It is within the realm of possibility for LSTM to bridge minimum delays of over 1000 discrete-time steps without sacrificing short-time delay capabilities by impelling a steadfast error flow through a Constant Error Carousel (CEC) within a particular unit. The processing and prediction of time series data can be effectively handled by using LSTM in the study of DFER, which effectively handles the temporal dependency of facial expression changes. At the same time, since LSTM has the effect of improving the robustness of the model to noise and uncertainty, LSTM can also effectively cope with the challenges brought by illumination changes and facial occlusion to DFER. After that, LSTM can be expanded to Bilateral LSTM (Bi-LSTM) [9]. Prediction can be achieved through the use of both past and future information by Bi-LSTM, but it has higher computational complexity and memory requirements than unilateral LSTM. Recently, a study proposed a method to use 3D-CNN and LSTM to extract the

DOI: 10.54254/2755-2721/87/20241545

provisional relationship between consecutive frames in video sequences, and found that the facial recognition rate was improved to a certain extent [10].

3. Discussion

In daily life, facial expressions are not static but dynamic, which makes video-based facial expression recognition became a mainstream trend. Although traditional models and deep learning models can roughly finish facial expression recognition, they are still facing many limitations and challenges. Conventional models have limited feature representation capabilities and rely on manual feature extraction, which may lead to an inaccurate capture of tiny shifts in facial expressions and longitudinal data present in dynamic facial expression sequences. While spatial feature dimensionality has been reduced via the use of Principal Component Analysis (PCA) [11], their ability to address the complexity of dynamic features is still limited. Therefore, evolving towards more complex models is necessary.

Deep learning models leverage large datasets and complex layer structures to learn rich feature representations, which are more conducive to research in DFER compared to traditional models. However, they also introduce new challenges. For instance, their internal mechanisms are more complex, making the decision-making process difficult to interpret, thereby reducing model interpretability and credibility. While the use of Grad-CAM improves model interpretability and transparency [12], the black-box spontaneous of deep learning models remains a significant obstacle in sensitive areas, which is widely accepted.

There are still many issues worth exploring in the topic of DFER research. For example, the processing time for single-frame data is lengthy, and the efficiency of storing and managing large amounts of video data is low. In the future, the Apache Spark framework may serve as an effective tool for distributed management and processing of large datasets [13]. External factors such as lighting changes, occlusion, and different shooting angles also pose challenges because they can significantly affect feature recognition accuracy. Developing more robust models that generalize well under different conditions is crucial. Furthermore, diversity in facial expressions due to factors such as gender, age, and ethnicity is often lacking in current datasets, restricting the practical application of models in real-world scenarios. Therefore, creating ideal dynamic facial expression datasets that include more attributes like gender, age, and ethnicity is necessary. Currently, only combining Transfer Learning (TL) can alleviate the problem of data imbalance and scarcity [14, 15]. Face video data may also contain sensitive information such as personal identity, behaviour patterns, and emotional states, raising serious ethical and privacy concerns regarding data collection and usage. Ensuring data anonymity and secure storage is crucial to prevent misuse of information and protect personal privacy. Facial expressions are just one component of human expression behaviour in reality. The emergence of social media and numerous digital platforms emphasizes the importance of Multimodal Sentiment Analysis (MSA) methods [16] that analyze human opinions on something from text, audio, images, etc.

Future research should focus on developing more efficient and interpretable deep learning models, improving robustness to external variations, and optimizing frameworks for processing large-scale video data. Additionally, diversifying datasets, integrating multimodal emotional data, and creating more accurate and reliable DFER systems are essential. Safeguarding the privacy of adopted facial data is also crucial while facilitating the smooth progress of this research.

By addressing these challenges, DFER technology can make significant strides in more practical and widespread applications, ultimately advancing fields such as human-computer interaction, psychological health monitoring, and public safety.

4. Conclusion

This work conducts a comprehensive survey on traditional models and machine learning models used for DFER, with a focus on HOG, SVM, CNN, and LSTM algorithms. Even with the advancements made possible by the use of both traditional and deep learning models, a number of obstacles still exist. These include limitations in expressing dynamic facial features, sensitivity to external factors, limited data storage, and inefficient processing. Future research should focus on building datasets that include

DOI: 10.54254/2755-2721/87/20241545

diverse attributes such as gender, age, and race. Developing more robust recognition models to address the effects of occlusion, lighting, and angle variations on DFER is also essential. Ensuring the privacy and security of video data used in research is crucial for protecting personal identity and sensitive information. Utilizing multimodal emotion analysis models, which combine data from text, audio, and images, can enhance the practicality of facial expression recognition systems in real-world applications. Efforts to tackle these challenges can enhance the accuracy, reliability, and applicability of DFER in various domains.

References

[1] Jung H Lee S Yim J Park S & Kim J 2015 Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Santiago: IEEE) pp 2983–2991

[2] Kim S An G H & Kang S -J 2017 Facial expression recognition system using machine learning Proceedings of the International SoC Design Conference (ISOCC) (Seoul: IEEE) pp 266–267

[3] Hinton G E & Salakhutdinov R R 2006 Reducing the dimensionality of data with neural networks Science vol 313 (Washington D.C.: American Association for the Advancement of Science) pp 504-507

[4] Carcagnì P Del Coco M Leo M et al. 2015 Facial expression recognition and histograms of oriented gradients: a comprehensive study SpringerPlus vol 4 (Berlin: Springer) p 645

[5] Chandra M A & Bedi S S 2021 Survey on SVM and their application in image classification International Journal of Information Technology vol 13 (Mumbai: Bharati Vidyapeeth) p 1-11

[6] Pan X 2020 Fusing HOG and convolutional neural network spatial–temporal features for video-based facial expression recognition IET Image Processing vol 14 (London: Institution of Engineering and Technology) p 176-182

[7] Mao L Chen S & Yang D 2021 Guided convolutional neural network video pedestrian action classification improvement method Journal of Wuhan University (Information Science Edition) vol 46 (Wuhan: Wuhan University Press) p 1241-1246

[8] Bai M & Goecke R 2020 Investigating LSTM for micro-expression recognition Companion Publication of the International Conference on Multimodal Interaction pp 7-11

[9] Siami-Namini S Tavakoli N & Namin A S 2019 The Performance of LSTM and BiLSTM in Forecasting Time Series Proceedings of the IEEE International Conference on Big Data (Big Data) (Los Angeles: IEEE) pp 3285-3292

[10] Hasani B & Mahoor M H 2017 Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (Honolulu: IEEE) pp 2278-2288

[11] Kang J Lin X & Wu X 2015 Face recognition algorithm based on Laplace pyramid dimension reduction Journal of Shaanxi University of Science and Technology (Natural Science Edition) vol 35 (Xi'an: Shaanxi University of Science and Technology) p 165-168174

[12] Selvaraju R R et al. 2017 Grad-cam: Visual explanations from deep networks via gradient-based localization Proceedings of the IEEE international conference on computer vision (Venice: IEEE)

[13] Li H Tan Y & Wu F 2017 Massive video face extraction and recognition parallel framework design and optimization Computer Applied Research vol 34 (Beijing: Science and Technology of China Press) pp 3811-3815

[14] Liang Z Liu D & Sun Y 2022 Micro-expression recognition method combining migration learning and separable three-dimensional convolution Computer Engineering vol 48 (Beijing: Science Press) pp 228-235

[15] Qiu Y Hui Y Zhao P Wang M Guo S Dai B Dou J Bhattacharya S & Yu J 2024 The employment of domain adaptation strategy for improving the applicability of neural network-based coke quality prediction for smart cokemaking process Fuel vol 372 (Amsterdam: Elsevier) p 132162

DOI: 10.54254/2755-2721/87/20241545

[16] Yan J Lu G Li H & Wang S 2018 Dual-modal emotion recognition based on face expression and speech Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition) vol 38 p 60-65

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241588

Analysis of traffic accidents based on Spark and causal inference

Quanjin liu

School of Computer Science, Wuhan University

2019302110266@whu.edu.cn

Abstract. Traffic accidents have occurred frequently in recent years, causing great losses to personal and property safety. Studying traffic accident data is helpful to identify key factors of traffic accidents from big data. This paper processes and calculates big data based on the Spark platform. By introducing causal inference into the analysis of traffic accidents, it establishes the causal relationships between 17 factors and the severity of traffic accidents, thereby analyzing the root causes and intermediate causes of the accidents. In addition, this paper also conducts an intervention study to evaluate the influence weight of each factor. The study finds that the physical conditions of pedestrians and weather conditions are inferred to be the root causes, and the others are intermediate causes. Besides that, the presence of police force and reduced traffic volume are considered to be the best ways to reduce traffic accidents. Therefore, this article believes that in real life, we should reduce the incidence of traffic accidents by controlling traffic flow and increasing patrol or on-duty police force. These findings provide a scientific basis for traffic management departments to develop more effective traffic safety strategies.

Keywords: traffic accident, big data, spark, causal inference, intervention study

1. Introduction

1.1. Research Background

Over the past century, cities have been continuously developing. As a result, roads have become increasingly crisscrossed, and cars have gradually become a necessary part of citizens' daily lives. At present, a complicated transportation system has been established, with private cars being seen everywhere. However, this phenomenon leads to traffic congestion and frequent accidents, consequently resulting in huge personal and property losses. The economic losses caused by traffic accidents are about 30 billion dollars each year, equivalent to approximately 3% of the gross national product of countries around the world[1].Therefore, how to effectively reduce the occurrence of traffic accidents, how to respond quickly after a traffic accident occurs, and how to reduce losses in an accident have become significant issues.

1.2. literature review

In recent years, researchers have widely utilized big data to explore traffic accident data to find reasons and suggestions. This technology can provide new solutions efficiently in terms of exploring the causes behind accidents. In addition, it can help enrich existing accident analysis methods and avoid analytical

DOI: 10.54254/2755-2721/87/20241588

bias due to the subjective experience of analysts[2].Many scholars have successfully analyzed the several causes of accidents using big data and have a certain degree of confidence in predicting occurrence of accidents. Luo Yulin built an accident diagnosis model by combining random forest, decision tree and generalized regression neural network, quantitatively estimated the influence of each factor, and output the severity rating of the accident. Finally, the latter had a higher prediction success rate and was supported by examples[2]. Zheng Lai et al. established a T-S fuzzy fault tree with major traffic accidents as the top event, people, vehicles, roads and environment as the intermediate events, and 24 sub-factors as the basic events. They converted it into a Bayesian network, and then bidirectionally inferred the importance and posterior probability of the basic events to determine the main causes. The accuracy and reliability of the cause analysis results of major traffic accidents were improved through forward and reverse reasoning[3]. Han Tianyuan et al. constructed a hierarchical model of the mechanism of serious and major road traffic accidents based on text mining. The results show that the contribution values of the network of causes of major road traffic accidents are illegal behavior, safety hazards and improper operation from large to small. The coupling of the direct causes of illegal behavior and improper operation and the indirect causes of safety hazards is the fundamental reason for the instability of the safety operation system of major accidents[4].

With regard to a certain technology of big data, spark is especially suitable for the transportation field because of its memory-based computing characteristics. Spark makes it fast in processing distributed data which is exactly the type of traffic accident data. Many researchers have used Spark to conduct research in the transportation field. Ebtesam Alomari et al. proposed a method to automatically detect road traffic-related events from Saudi dialect tweets using machine learning and big data[5]. A. Saraswathi et al. implemented a real-time traffic monitoring system based on Spark. In the article, the traffic volume is predicted using connected vehicles and real-time streaming data is processed using Apache spark and the traffic volume is displayed on a dashboard using springboot[6]. Guo Yuda et al. designed and implemented an efficient parallel algorithm based on the Spark computing framework for road network segmentation and kernel density calculation in road network kernel density estimation. Taking traffic accidents as an example, four groups of experiments were conducted for comparative analysis. The results show that the road network kernel density estimation parallel algorithm based on the Spark computing framework has high computational efficiency and good scalability[7].Spark can help process a large amount of traffic accident data in this experiment. This paper attempts to study traffic accidents based on causal inference on the Spark platform.

1.3. Research gaps

Admittedly, existing research has certainly analyzed the impact of various factors on traffic accidents and can obtain the weight ranking of each factor. However, most of the previous studies used traditional methods such as neural networks and regression analysis, which lacked in-depth discussion of causal relationships, making it difficult to identify the true causes of accidents among various factors. Meanwhile, many studies focused on the relationship between a single factor and traffic accidents, but the data dimension is limited and insufficient to reflect the overall situation. Furthermore, most or a large portion of current research utilized limited data samples, which may fail to accurately represent the overall situation. Existing research has not explored the potential for improvement through targeted interventions, which is a significant gap in the current literature. If causal inference is introduced to the research in traffic accident data and Bayesian networks (BN) are used to introduce external interventions, it can help define the causal effects of external interventions and describe the causal relationships between multiple variables related to accidents[8]. Consequently, it helps obtain conclusions on how much accident reduction can be achieved under certain intervention, thus effectively identifying the improvement priorities in accident prevention practice and guiding relevant activities.

Causal inference is a cutting-edge direction currently, and has not yet been applied in the field of traffic accident analysis and prediction. However, it has achieved success in fields such as medicine and education. Zhang Yu solved the self-selection bias in the field of education based on causal inference[9]. Li Shiyuan et al. studied the causal relationship between the participation of extracurricular tutoring and

DOI: 10.54254/2755-2721/87/20241588

the generation of negative emotions of middle school students in mainland China based on two phases of data from the China Education Longitudinal Survey, and answered questions such as "whether participation in extracurricular tutoring causes depression" and "who is depressed" at a quantitative level[10]. Liu Xinhui et al. started with realistic data and used series of causal inference methods to screen health index indicators that have evidence-based causal relationships with health/disease outcomes, which can provide more practical and valuable real-world evidence for health/disease management[11].

1.4. Research Topic and method

In this study, the research steps are divided into two distinct components: the data processing phase, which involves preparing the data for analysis, and the causal inference phase, which involves identifying the causal relationships between variables.

This experiment uses the traffic accident data of the United States from 2005 to 2007 as the data set, and uses Pandas to process the relevant data. This data set has 33 dimensions. First, the irrelevant data of the traffic accident link is predicted to be cleared, and then the missing values will be cleared or filled. Afterwards, the object data type is indicated to be normalized, and finally a data set that can identify causal relationships is obtained.

After the above processing, the CasualModel of the Dowhy library is introduced for causal relationship identification, and Matplotlib is used for visualization to obtain the causal relationship diagram of the entire data set. Through these operations, which factor should be the cause, which factor ought to be the effect and the relationship of them could be revealed.

After analyzing the causal relationship from the obtained causal relationship diagram, this article will continue to discover the results of interventions. The variables are estimated in the obtained causal relationship to obtain the influence weight of each factor, thereby evaluating their intervention effects. Finally, in order to evaluate the accuracy and authenticity of the experiment, a refutation test is performed.

2. Experiment

2.1. data processing

This experiment utilizes a comprehensive dataset of traffic accidents in the United States, spanning from 2005 to 2007, which serves as the foundation for our analysis. This data set has many dimensions, covers a comprehensive range, and has a huge amount of data. The specific data types are shown in table 1.

Table 1. Unprocessed data types

Accident_Index	object	Location_Easting_OSGR	float
Accident_Severity	int	Location_Northing_OSGR	float
Number_of_Vehicles	int	Did_Police_Officer_Attend_Scene_of_Accident	object
Number_of_Casualties	int	Local_Authority_(Highway)	object
Light_Conditions	object	Pedestrian_Crossing-Human_Control	object
Weather_Conditions	object	Pedestrian_Crossing-Physical_Facilities	object
Carriageway_Hazards	object	Road_Surface_Conditions	object
Urban_or_Rural_Area	int	Special_Conditions_at_Site	object
Local_Authority_(District)	int	LSOA_of_Accident_Location	object

DOI: 10.54254/2755-2721/87/20241588

Table 1. (continued).

Longitude	float	Police_Force	int
Latitude	float	Day_of_Week	int
Date	object	Road_Type	object
Time	object	Speed_limit	int
Year	int	Junction_Detail	float
Junction_Control	object	1st_Road_Number	int
1st_Road_Class	int	2nd_Road_Number	int
2nd_Road_Class	int		

From the above table, the data dimension statistics are composed of 5 items of float type, 13 items of int type, and 15 items of object type. This data set includes factors, some of which are considered to be the causes affecting traffic accidents, such as human factors, pedestrians, environmental factors, weather, etc. Some other factors are traffic accident severity evaluation criteria, such as accident severity rating and number of injured victims. The data set contains a total of 570,000 traffic accident data. Therefore, the data set can meet the experimental requirements, and the final conclusion is also of reference value.

There are some irrelevant variables in this data set, such as the accident ID and the year of the accident, which are not very useful for studying the cause of the accident. Apart from these data, missing values are also a large part of the data set and need to be processed as well. This article uses the drop() method and dropna() method of Python's Pandas library to process the irrelevant variables and missing values.

Afterwards, since the Dowhy library used for causal inference cannot process the object type, an effective method must be used to convert object type data in the data set into integer or float type. The LabelEncoder method of the sklearn.preprocessing library can help. It uses a value between 0 and number of categories - 1 to encode the target label and convert non-numeric data into numeric data. This paper processes 15 object type data by this method. After the above series of data processing, the final data set consists of 18 dimensions and 315,000 traffic accident data. The processed data types are shown in table 2.

Table 2. Processed data types

Day_of_Week	int	Police_Force	int
Accident_Severity	int	Time	float
Number_of_Vehicles	int	Did_Police_Officer_Attend_Scene_of_Accident	int
Number_of_Casualties	int	Speed_limit	int
Light_Conditions	int	Pedestrian_Crossing-Human_Control	int
Weather_Conditions	int	Pedestrian_Crossing-Physical_Facilities	int
Carriageway_Hazards	int	Road_Surface_Conditions	int

DOI: 10.54254/2755-2721/87/20241588

Table 2. (continued).

Urban_or_Rural_Area	int	Special_Conditions_at_Site	int
Road_Type	int	Junction_Control	int

2.2. causal relationships identification

Actually, there must be certain causal relationships between the variables that cause traffic accidents. However, it is difficult to construct a correct causal relationship graph based solely on prior knowledge. The causal inference library(Dowhy) can help with causal relationship discovery. PC and GES are widely used and fast causal discovery methods. Therefore, this paper attempts to discover the causal relationship graph by introducing the PC and GES methods, and visualize the causal graph through the GraphUtils library and the pyplot library.

There is a significant difference between the causal diagrams obtained by PC and GES methods. The PC algorithm was proposed by Peter Spirtes et al. It is an algorithm based on conditional independence testing. The core idea of the PC algorithm is to construct a causal graph by observing the variable pairs in the data to determine whether they are directly or indirectly related based on conditional independence tests. On the other hand, the GES algorithm assumes that there is no causal relationship between most variables and tries to find a sparse causal structure. When analyzing the relationship between the severity of traffic accidents and light conditions, GES believes that the severity of traffic accidents causes light conditions. Thus, comparing the causal graphs obtained by PC and GES, PC is more in line with prior knowledge and more correspond to the reality. However, according to prior knowledge, light conditions are objective conditions, and the causal relationship between the two should be that, potentially, light conditions cause traffic accidents to a certain extent. So there is a contradiction between the GES causal diagram and prior knowledge. While, the causal relationship analyzed by PC is consistent with prior knowledge. This paper analyzes the causal relationship obtained by the PC method.

The causal graph clearly illustrates 2 nodes, Pedestrian_Crossing-Physical_Facilities and Weather_Conditions, as the initial nodes, which are the root causes. The police force involved in the accident handling is Police_Force, and the accident severity is Accident_Severity, which are the final nodes and the final results of the entire cause-effect diagram.

The accident severity is directly affected by Special_Conditions_at_Site,Junction_Control, Number_of_Vehicles,Light_Conditions,Speed_limit,Number_of_Casualties,Urban_or_Rural_Area,Did_Police_Officer_Attend_Scene_of_Accident,Weather_Conditions and other variables, and is indirectly affected by other variables. It does prove that traffic accidents are the result of multiple factors.

The node Police_force is directly affected by some variables such as Pedestrian_Crossing-Physical_Facilities,Junction_Control,Road_Surface_Conditions,Urban_or_Rural_Area,Did_Police_Officer_Attend_Scene_of_Accident, and is also indirectly affected by some other variables. These relationships reveal that when the factors which result in traffic accidents are different, the police force dispatched will change accordingly.

Besides, the number of casualties, a variable that people pay attention to, is related to Road_Surface_Conditions,Speed_limit,Did_Police_Officer_Attend_Scene_of_Accident,Light_Conditions,Junction_Control,Number_of_Vehicles,Urban_or_Rural_Area. It shows that whether an accident will cause casualties is potentially related to the condition of the road surface, the speed limit, the presence of police, the lighting conditions, whether the traffic lights at the intersection are working properly, the volume of traffic, and whether it is in an urban area.

2.3. Estimation

If changing a variable lead to a change in the final dependent variable, then it can be defined that the variable will cause the dependent variable to occur. In this process, everything else remains unchanged.

DOI: 10.54254/2755-2721/87/20241588

Therefore, in this step, this article identifies the causal relationship and effect to be estimated through the properties of the causal graph.

In this paper, all factors except Accident_Severity are estimated as intervention parameters. First, set the parameters treatment and outcome to the variable to be estimated and the dependent variable respectively, and establish a causal model for the variable to be estimated and the accident severity (Accident_Severity). Then utilize the identify_effect() method and estimate_effect() method of the causal model for estimation. This paper estimates the intervention of each factor through the backdoor criterion of structural causal model.

In the estimation step, the estimated values of other variables are shown in figure 1

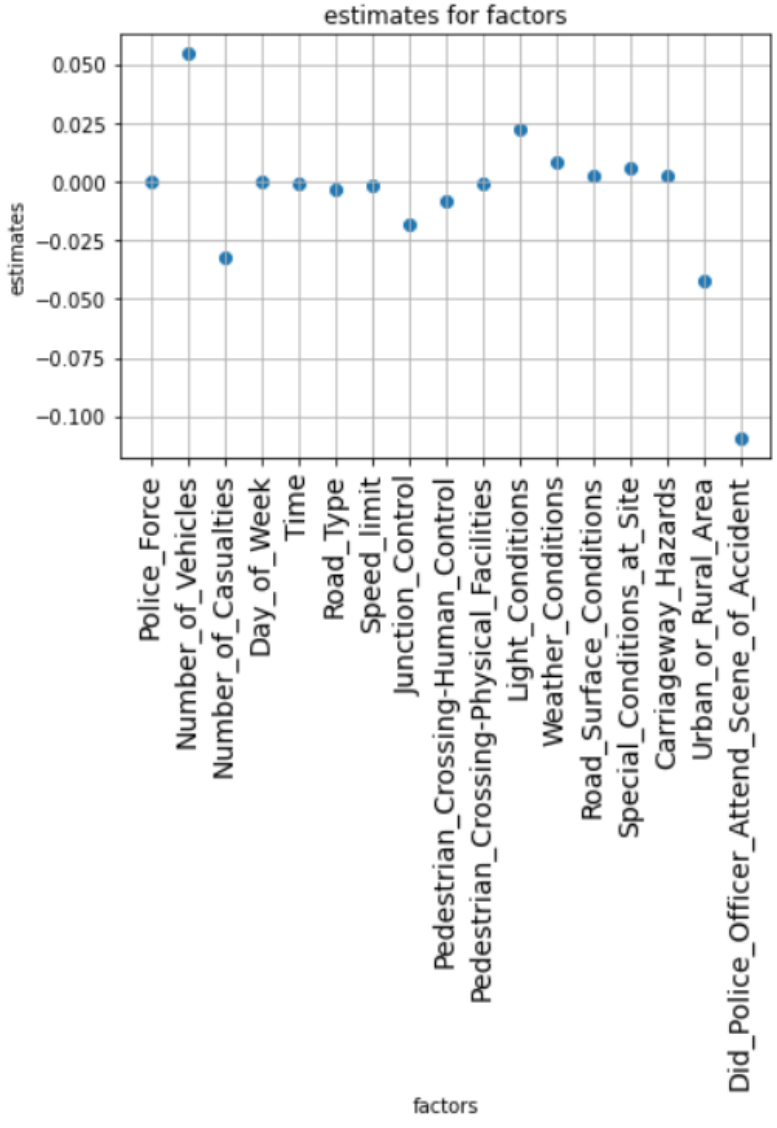

Figure 1. Estimated values.

Number_of_Vehicles,Number_of_Casualties,Did_Police_Officer_Attend_Scene_of_Accident,Urban_or_Rural_Area and Light_Conditions have significantly greater estimates of accident severity than other variables. It shows that the size of the traffic volume, the number of casualties, whether there are police on the scene, whether it is a urban area or a rural area, and the quality of lighting conditions have

DOI: 10.54254/2755-2721/87/20241588

a significant causal relationship with the severity of traffic accidents. Other factors need to have an impact by acting on these five factors.

2.4. Refutation Test

In order to verify the reliability of the above causal estimation results, we conducted a refutation test. In fact, the causal relationship in the causal diagram obtained from the second step is only a hypothesis about various factors in this article. The refutation test is to verify whether this hypothesis is correct. The basic idea of the refutation test is using the random common cause method. Random common causes are random, non-specific causes that are prevalent in the observed data and affect multiple variables. These causes are often unpredictable and interfere with inferences about causal relationships. In this article, this method is used to eliminate the influence of these interferences. Its specific operation is adding random covariates to the data and rerun the analysis to see if the causal estimate changes. If the hypothesis is correct at the beginning, the causal estimate should not change much.

This paper uses statistical tools to evaluate the interference caused by this random variable. Among them, the P value is a key concept. P value is a parameter in statistics used to evaluate the significance of the difference between observed data and the null hypothesis. It represents the probability of observing the current sample or more extreme cases if the hypothesis is true. Assuming that this random common cause has no relationship with the causal relationship, after the refutation of this method, the obtained P values should be relatively large, thereby lending support to the original causal hypothesis.

However, it is important to note that a large p-value only indicates that the added random common cause has little influence on the causal estimate, but it cannot directly prove that the original hypothesis is correct, as there may be other unconsidered influencing factors.

After using the random common cause method, the new effect value is shown in figure 2, and the ratio of the original effect to the new effect is shown in figure 3.

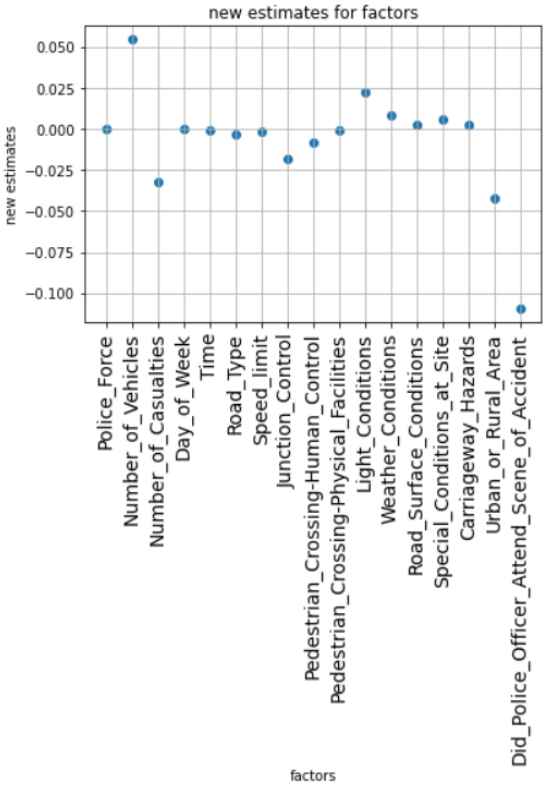

Figure 2. Re-estimated values.

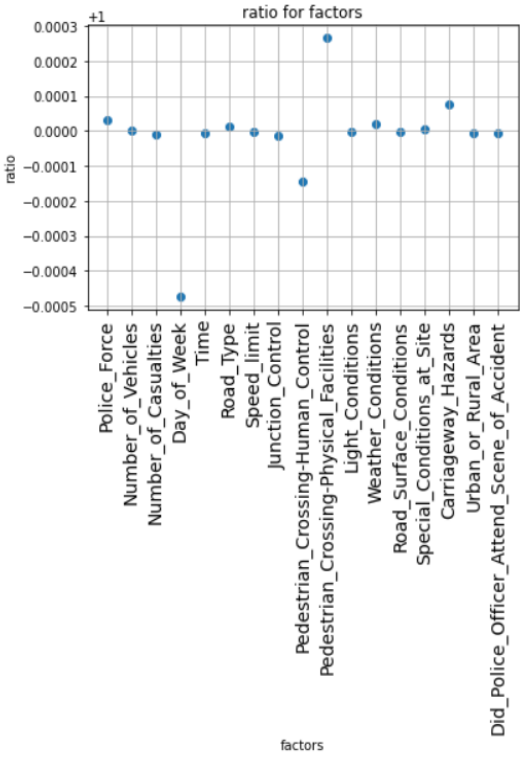

Figure 3. Ratio values.

DOI: 10.54254/2755-2721/87/20241588

The ratio of the original effect to the new effect is generally close to 1, which indicates that the causal estimate does not change much. It shows that the original hypothesis is correct.

The P value is shown in figure 4.

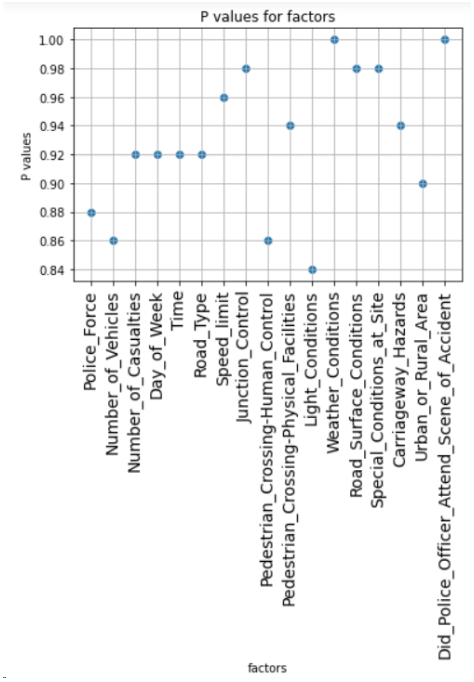

Figure 4. P values.

As shown in the data in the figure, the P values are generally high, and the P values of the two variables Did_Police_Officer_Attend_Scene_of_Accident and Weather_Conditions reach 1. This suggests that the addition of random common causes has almost no effect on the cause-effect relationships. Therefore, the causal relationships in the causal graph obtained in the second step are hardly affected by the covariates. Thus, it reflects that the credibility of the causal relationship between these seventeen factors and Accident_Severity is very high.

3. Conclusion

This study identifies the complex causal relationships between all factors, among which the physical conditions for pedestrian traffic and weather conditions are the root causes that lead to traffic accidents and the dispatch of police forces. In the estimation, this paper finds that intervening with traffic volume and presence of police has the large impacts on the severity of traffic accidents. In addition, this paper conducts a refutation test using the random common cause method and statistical methods, and the results shows that the above causal relationship is highly credible. Therefore, the results of the causal relationships mining and intervention experiment are of reference value for the actual situation. Controlling traffic flow and increasing police patrols can effectively help reduce the occurrence of traffic accidents.

This paper introduces causal inference into the field of traffic accident analysis for the first time. It not only conducts a causal relationship study to reveal the reasons that lead to the severity of the accident, but also, from the perspective of improvement, does research. This can provide future researchers with a new method to study traffic accidents. By analyzing the improvement of traffic accidents from the perspective of intervention, this study helps traffic management departments to identify improvement priorities and reduce the occurrence of traffic accidents.

This study relies on causal inference and only identifies the overall causal relationship of all factors. In the future, the specific causal relationships between each factor can also be studied, so as to show the

DOI: 10.54254/2755-2721/87/20241588

causal relationships more deeply and specifically, and to use this to determine the root cause of the effect of the intervention.

References

[1] X.G. Guo (2020). Research on prediction of road traffic accident severity based on Spark platform. Yunnan University.

[2] Y.L. Luo (2020). Highway traffic risk prediction based on big data. Changsha University of Science and Technology.

[3] L. Zheng, P. Gu and J. Lu (2021). A cause analysis of extraordinarily severe traffic crashes based on t-s fuzzy fault tree and bayesian network. Journal of Transport Information and Safety, 39(4): 43-51+59.

[4] T.Y. Han, S. Tian, K.G. Lyu, X. Li, J.T. Zhang and L. Wei (2021). Network analysis on causes for serious traffic accidents based on text mining. China Safety Science Journal, 31(9): 150-6.

[5] E. Alomari, R. Mehmood and I. Katib (2019). Road traffic event detection using twitter data, machine learning, and apache spark, 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), Leicester, UK, pp. 1888-95.

[6] A. Saraswathi, A. Mummoorthy, A. Raman G.R. and K. P. Porkodi (2019). Real-time traffic monitoring system using spark, 2019 Int. Conf. on Emerging Trends in Science and Engineering (ICESE), Hyderabad, India, pp. 1-6.

[7] Y.D. Guo, X.Y. Zhu, W. Guo and B. She (2020). Parallel algorithm for road network kernel density estimation based on spark computing framework. Geomatics and Information Science of Wuhan University, 45(2): 289-95.

[8] Z.G. Ma, X.H. Xu and X.E. Liu (2022). Three analytical frameworks of causal inference and their applications. Chinese Journal of Engineering, 44(7): 1231-43.

[9] Y. Zhang (2013). Causal inference model in quantitative evaluation of education policy and the implications of mixed methods. Tsinghua University Education Research, 2013(3):29-40.

[10] S.Y. Li and A.Y. Liu (2022). “Melancholic children”: does cram school participation lead to negative emotions? causal inference based on chinese education panel survey (CEPS) data. Chinese Journal of Sociology, 42(2): 60-93.

[11] X.H. Liu, H.K. Li, L.J. Wang, A.L. Liu, Y. Qi, S.S. Sun, L.F. Zhang, H.J. Ji, G.Y. Liu, H. Zhao, Y.N. Jiang, J.Y. Li, C.C. Song, X. Yu, L. Yang, J.C. Yu, H. Feng, F.J. Yang and F.Z. Xue (2022). Causal inference methodology for the screening of indicators for health indices. CHINESE JOURNAL OF DISEASE CONTROL & PREVENTION, 26(10): 1180-6.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241537

Exploration of the application of artificial intelligence in modern agricultural production—Take orchard management as an example

Miaowei Wang

Shandong University of Finance and Economics, Shandong, 271100, China

1479039715@qq.com

Abstract. In the context of increasing global agricultural challenges, the application of artificial intelligence technology in the agricultural field is increasingly a trend, especially in the production of agricultural products, intelligent identification technology has shown the potential to significantly improve production efficiency and optimize yield and quality. By the mid-21st century, demand for food production is expected to reach 50 percent, and there will be enormous pressure to achieve this goal with traditional agricultural technologies, which could be achieved through the application of artificial intelligence. The application of artificial intelligence technology in modern agricultural production will be analyzed in detail in this paper, the guidance of future research fields will be proposed, and the existing challenges and technical problems will be identified and discussed in order to promote the deepening and wide application of intelligent agriculture. This paper specifically discusses examples of applications in orchard management, pest detection, and automated harvesting and summarizes the effectiveness and obstacles of these techniques.

Keywords: Agriculture Intelligence, artificial intelligence, intelligent identification technology, orchard management

1. Introduction

Technology-intensive agricultural production mode has gradually replaced the traditional labor-intensive mode, which is particularly critical in agriculture, the basic industry of human society. In recent years, the progress of deep learning and computer vision technology has significantly promoted the application of artificial intelligence in agricultural production. As a major apple producer, China needs to consume a lot of manpower and material resources in its picking stage. Therefore, China gradually began to develop intelligent agriculture. In the fields of fruit and vegetable cultivation management, disease and pest warning, and agricultural harvesting automation, advanced identification technologies, such as YOLOv5-driven target detection algorithms, have shown significant potential and efficiency [1]. The modern agricultural landscape is witnessing an unprecedented foray of Artificial Intelligence (AI) technology, which is profoundly transforming productivity and output quality through sophisticated smart recognition mechanisms. Focusing particularly on the intricate dynamics of orchard management, we uncover AI's exceptional capabilities in pinpointing diseases and pests with accuracy, and its transformative influence on the sphere of automated harvesting. Within the domain of crop yield, the

DOI: 10.54254/2755-2721/87/20241537

assimilation of AI has unleashed a significant surge in operational efficiency, lessened dependence on manual labor, and fostered superior-grade produce, thereby escalating their competitive edge in the market. Bridging the gap between practical agricultural situations and extensive scholarly research, this investigation sheds light on the myriad uses and technological prowess of AI in farming, ultimately striving to propel the growth of China's agricultural sector [2].

2. Overview of intelligent agricultural technology applications

The infusion of artificial intelligence into the fertile soil of agricultural methods is sowing the seeds of a groundbreaking revolution, fundamentally reforming the way we nurture and reap the bounties of our earth. The labor-intensive conventional farming approach frequently struggles with inefficiencies and precision deficits. Thanks to cutting-edge artificial intelligence technologies, including big data analysis, computer vision and image recognition technologies, automated and intelligent agricultural production has been realized. Fruit tree cultivation and management is a key stage of agricultural production, covering many fields such as planting technology, pruning process, irrigation and fertilizer application as shown in figure 1. For example, Liu Zilong et al. proposed the upgraded YOLOv5 algorithm, combined with the coordinate attention mechanism, perceptron components and adaptive spatial feature fusion strategy, which can significantly improve the efficiency of apple growth monitoring [1]. Intelligent irrigation and fertilization systems, based on artificial intelligence technology, accurately evaluate the water and nutrient needs of crops and realize accurate irrigation and fertilization to reduce resource consumption. By integrating artificial intelligence technology into orchard operations, not only can the environmental burden be reduced, but also the management efficiency can be significantly improved, thus promoting the development of green and sustainable agriculture.

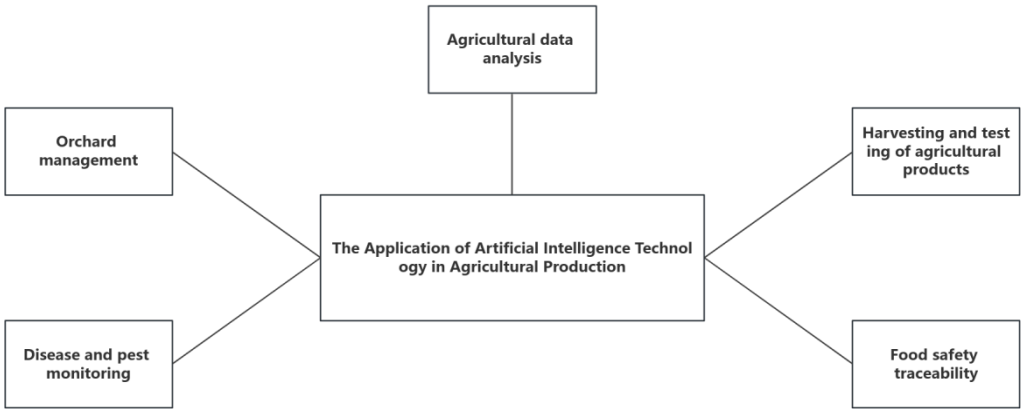

Figure 1. The relationship diagram of AI application in agriculture

3. Specific application of artificial intelligence in orchard management

3.1. Application in the monitoring of diseases and insect pests

In the process of crop production, the pests and diseases always constitute a major agricultural challenge. The occurrence of major crop diseases and insect pests may lead to a significant decline in yield and cause irreversible damage to national agriculture and peasant household economy. Real-time crop health monitoring can be achieved by using image recognition technology and using deep learning methods to achieve pest control. With the help of remote sensing technology and Internet means, the detection of diseases and pests of crop plants can be implemented, which can accurately identify diseases and pests. For example, Shi Jialu et al. proposed that through image acquisition technology to analyze the plants affected by diseases and pests, we can identify the specific pest category[3]. Through the interaction of image recognition technology and the Internet, these functions can be realized; on the other hand, convolutional neural network (CNN) can also play a key role by learning the images of a large number

DOI: 10.54254/2755-2721/87/20241537

of diseases and pests, automatically extract and identify the unique characteristics of diseases and pests, improve the accuracy of diseases and pest diagnosis, enable farmers to take effective control measures and reduce pesticide dependence.

3.2. Application in agricultural product harvest

As the amber glow of fall blankets the countryside, farmers delve into the strenuous endeavor of harvesting nature's abundance.For ages, Chinese agriculture has been steeped in labor-intensive techniques and rudimentary tools, a centuries-old tradition that now stands in stark contrast to the rapid advancements in artificial intelligence. The drawbacks of these conventional approaches have increasingly become evident, primarily their time-consuming nature, physical rigor, and, sadly, the risk of crop wastage and inefficient resource utilization. However, the emergence of AI technology signals a paradigm shift, ushering in an era of smart and autonomous farming. This revolutionary innovation not only amplifies crop yields exponentially but also fortifies the financial robustness and competitive stance of the agricultural sector. A groundbreaking revolution lies in an innovative visual system, skilled in decoding the intricate layers of crop canopies. This paves the way for a deeper comprehension of the spatial dynamics that unfold within fruit orchards, ensuring the seamless and optimized operation of robotic harvesters. The secret to unlocking the latent power of our farming techniques resides in this trailblazing technology [4]. As proposed by Gao Rongliang, integrating image recognition with the might of big data can create algorithmic models that mimic and elevate existing harvesting approaches. By harnessing the power of data analytics, it refines operational routes and tactics, thus amplifying productivity. Agricultural robots, armed with intelligent recognition technology, can now accurately identify ripened fruits and execute their picking duties with remarkable precision, thereby significantly boosting harvest quantities and reliability [5]. By leveraging advanced artificial intelligence systems and fusing them with cutting-edge camera and sensor technologies, the exact pinpointing of fruits amidst complex agricultural terrains becomes achievable. As a result, a dexterous robotic arm performs the harvesting task with a delicate touch, ensuring minimal fruit bruising and maximum crop yield. This innovative approach caters to the escalating need for streamlined and efficient agricultural output, appealing to a broader audience concerned with sustainable food production.

3.3. Application in agricultural product detection

As scientific advancements march forward, the bar for product excellence is consistently raised. In the realm of agricultural produce assessment, conventional approaches fall short due to their sluggish pace and susceptibility to human bias. By harnessing the power of image recognition technology and sophisticated machine learning algorithms, we can promptly and precisely evaluate both the external appearance and hidden quality of agricultural goods, revolutionizing the industry. Step-by-step classification benefits from the optimized image recognition algorithm. By detecting the color, shape, and size of fruits and vegetables, we efficiently screen high-quality products, improve the quality utilization rate of various products, and reduce loss.

3.4. Application in data analysis

Accurate agricultural data analysis plays a crucial role in shaping national agricultural development strategies. Leveraging artificial intelligence technology to optimize agricultural data analysis can effectively provide more precise and robust support for the country's agricultural strategies. In a recent study conducted by Ye Ting and her team in 2022, they explored the application of data mining in the agricultural sector. The research highlighted the significant potential of integrating data mining technology into agricultural production to improve the efficiency and scientific validity of agricultural management and operations[6].

By tapping into the capabilities of artificial intelligence, a profound comprehension of agriculture's past and present scenarios unfolds, shedding light on the essence of crop vitality, weather fluctuations, and soil attributes. This knowledge equips farmers with the tools to devise sharper and more efficient cultivation approaches. A predictive model, powered by the LSTM algorithm's sophisticated memory,

DOI: 10.54254/2755-2721/87/20241537

accurately forecasts optimal planting and harvesting moments, aligning with crop life cycles and meteorological data. Integrating cutting-edge recognition technology and extensive data analysis, the incessant well-being of crops is diligently monitored, facilitating immediate interventions and insightful choices.This pioneering approach empowers farmers with the skill to promptly detect and remedy any agricultural mishaps, thereby mitigating potential hazards and reducing losses significantly. As a result, it boosts the overall efficiency of farming endeavors. Furthermore, the groundbreaking fusion of Artificial Neural Networks (ANN) and Multi-Objective Genetic Algorithms (MOGA) revolutionizes the way energy consumption and greenhouse gas emissions in fruit farming are modeled, paving the way for their optimal management[7].

3.5. Application in the traceability system of agricultural products

At the core of agricultural produce consumption lies the pivotal concern of food security, a decisive factor that profoundly influences the well-being and existential peace of mind of consumers. Ensuring food safety through conventional means, such as rigorous laboratory assessments and painstaking visual inspections, often entail intricate procedures and lengthy durations. In this context, the groundbreaking integration of artificial intelligence technology into developing a smart traceability system promises to greatly enhance protective measures, thereby revolutionizing the entire domain of food safety.

By harnessing sophisticated intelligent recognition technology alongside the robust backbone of blockchain technology and profound data insights, a revolutionary Smart Traceability ecosystem can be forged for seamless and all-encompassing food safety oversight and authentication. This cutting-edge system knits together smart labeling and the magic of QR codes, weaving a comprehensive narrative of food's journey from inception to delivery. In the fertile landscape of agriculture, this technology takes center stage, seamlessly uploading details of provenance and cultivation into the unalterable blockchain ledger, thereby fortifying the reliability of agricultural information. Furthermore, stealthy sensors nestled within the packaging serve as diligent guardians, monitoring environmental conditions and product integrity throughout their voyage, ensuring that only the highest quality offerings reach the hands of consumers

4. Future prospects and existing challenges

Although in the process of intelligent agricultural production, artificial intelligence technology has shown great development, but its popularization and implementation are still facing significant obstacles. Yuan Shufang mentioned in her publication "Discussion on the Application of Artificial Intelligence Technology in Intelligent Agriculture Production" that in the complex agricultural ecological environment, the accuracy of intelligent identification may encounter challenges, and it is necessary to continue to optimize the algorithm and expand the data sample database[8]. The high cost of introducing a large number of advanced equipment into modern agriculture is also a big problem. Globally, small farms account for a third of the total, compared with 80 percent in China, putting a heavy burden on many small farmers. Implementing a collaborative farming paradigm and fostering a spirit of collective endeavor could dramatically reshape the economics of rural regions. Coupling this innovative approach with state-of-the-art machinery and technological advancements paints a promising picture. However, the hurdles of limited tech penetration and farmers' inadequate technical prowess call for intensified efforts in knowledge dissemination, training programs, and skill-enhancing seminars. These steps would facilitate the seamless integration of intelligent agricultural innovations. Furthermore, the pivotal role of supportive policies and industry backing in hastening the adoption of smart agri-tech solutions cannot be overlooked. The crucial role of government involvement is undeniable, especially when it comes to empowering small-scale farmers with state-of-the-art tools and technological know-how. Through sponsorships and financial assistance, governments foster a paradigm shift in agriculture, boosting overall productivity. This empowerment doesn't just energize farmers; it ignites a transformative agricultural revolution, opening up new horizons in the domain.

DOI: 10.54254/2755-2721/87/20241537

5. Conclusion

A comprehensive delve into the realm of intelligent agriculture reveals the far-reaching impact and pervasive presence of artificial intelligence within the fertile soil of farming. It has revolutionized conventional farming methodologies, uplifting productivity and accuracy to unparalleled levels, and skillfully harmonizing resource allocation for enhanced crop quality. This discussion uncovers the unique contributions of mammoth data analysis, intricate deep learning algorithms, perceptive computer vision, IoT-driven sensor networks, revolutionary robotics, unmanned aerial vehicles (UAVs), and cutting-edge remote sensing technology. These interconnected innovations form a robust scaffold, propelling the evolution of modern agriculture.

Despite the progressive infiltration of intelligent agricultural technology, certain limitations persist. Sensors and Internet of Things (IoT) devices, though prolific data generators, still call for enhancements in data precision and reliability. Likewise, the promise of deep learning and computer vision technologies is conditional upon the caliber of data sets and the refinement of algorithms. The realm of agricultural robotics encounters challenges in navigational exactitude and economic feasibility. Though Unmanned Aerial Vehicles (UAVs) and remote sensing technology offer comprehensive surveillance, their operational reach and monitoring intensity could be more robust.

In the realm of future exploration, several key areas warrant attention. Primarily, advancements in data acquisition and processing methodologies are crucial to guarantee the excellence and uniformity of data. Secondly, refining deep learning algorithms and computer vision techniques can significantly enhance their efficacy within the agricultural domain. Concurrently, it's imperative to lessen the financial burden associated with agricultural robotics while boosting their operational precision. In conclusion, the exploration of Unmanned Aerial Vehicles (UAVs) and remote sensing technology must push beyond conventional limits to forge advanced and resilient instruments for agricultural oversight and governance. The realm of intelligent agriculture ushers in a fresh era of eco-conscious farming progress, revealing a wealth of untapped potential. Despite current challenges, the relentless evolution and enhancement of these technologies augur a pivotal role for smart agriculture in the future. It stands ready to significantly influence global food security and foster the sustainable transformation of agricultural methods across the globe.

References

[1] Liu Zilong, Zhang Lei. Improving YOLOv5 detection of small target apples in natural environments [J / OL]. Journal of Systems Simulation, 1-15 [2024-05-30].

[2] Kou wei. Application of agricultural mechanization in urban modern agricultural production [J]. Agricultural Technology Service, 2017,34 (08): 149.

[3] Shi Jialu, Zhang Haixia. Application of information technology for fruit tree diseases and insect pests [J]. Agricultural Engineering Technology, 2023,43 (23): 31-32.

[4] L.G.D, Divya R, Piranav S, et al. Estimating depth from RGB images using deep-learning for robotic applications in apple orchards [J].Smart Agricultural Technology, 2023, 6 100345-.

[5] Gao Rongliang. Research on the application of artificial intelligence technology in modern agricultural machinery [J]. Modern Agricultural Machinery, 2024, (03): 121-124.

[6] Ye Ting, Ma Hongjuan, Lu Rui, et al. Application of artificial intelligence in smart agriculture—Data mining and machine learning are taken as an example [J]. Smart Agriculture Guide, 2022,2 (18): 27-29 + 32.

[7] Artificial Intelligence; Researchers from University of Malaya Detail Findings in Artificial Intelligence (Resource management in cropping systems using artificial intelligence techniques: a case study of orange orchards in north of Iran) [J]. Ecology Environment & Conservation, 2016

[8] Shu-fang yuan. On the application of artificial intelligence technology in intelligent Agriculture production [J]. Agricultural Engineering Technology, 2024,44 (05): 29-30.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241546

Research on the application of statistical methods based on big data in the medical and health field

Wenyan Yang

Shandong University of Science and Technology, Qingdao, China

15610552907@163.com

Abstract. In the current era of information, the rise of big data technology has become a key force driving the development of the medical and health field. Statistical methods, as core tools for data analysis, have demonstrated unique advantages in processing vast medical datasets. This paper comprehensively analyzes the various applications of big data statistical methods in the medical and health field, focusing particularly on their practical effects and potential value in disease prediction, epidemiological research, medical resource optimization, and personalized medical services. Through a review of relevant literature and in-depth discussions of multiple cases, this paper reveals how statistical methods play a crucial role in improving diagnostic accuracy and medical service efficiency. It also addresses the challenges faced in actual applications, such as data privacy protection and technical standardization.

Keywords: Big data, statistical methods, medical health, disease prediction, personalized medicine.

1. Introduction

With the rapid development of big data technology, the medical and health field is experiencing an unprecedented transformation. Big data not only provides the ability to handle complex health information but also brings profound challenges and innovative opportunities to traditional medical models. Statistics, as a science of dealing with data relationships and analyzing data structures, has now become a bridge for interpreting these complex medical data. Through precise data processing techniques, disease prediction has become more accurate, epidemic control more effective, resource allocation more reasonable, and immense potential has been shown in personalized medicine. Against the backdrop of global health governance, effectively utilizing statistical methods can not only enhance the quality and efficiency of medical care but also play a crucial role during public health crises. However, as technology rapidly advances, how to use these data responsibly while ensuring individual privacy, and how to enhance the accuracy and reliability of data analysis, remain significant challenges in the medical and health field.

2. Basic Concepts and Developments of Big Data and Statistical Methods

2.1. Definition and Characteristics of Big Data Technology

Big data technology refers to the techniques for processing large-scale, multi-type data collections at high speeds, allowing valuable information to be extracted from vast amounts of data. In the medical

DOI: 10.54254/2755-2721/87/20241546

and health field, these data often come from electronic health records, medical imaging, genomic data, patient-reported information, and the internet. The core characteristics of big data include large volume, a wide variety of types, rapid processing speed, and low value density, which make big data technology a key tool for transforming medical and health information [1].

2.2. Application of Statistical Methods in Big Data Analysis

The application of statistical methods in big data analysis in the medical field mainly focuses on data organization, analysis, and interpretation. This includes using descriptive statistics to summarize the basic characteristics of data, employing inferential statistics to generalize from sample data to larger populations, and utilizing predictive models and machine learning algorithms to predict future events or determine correlations between variables. For example, by applying regression analysis, researchers can understand and predict the effects of certain drugs on different patient groups, or identify patient groups with similar disease presentations through cluster analysis.

2.3. Development Trends and Technological Innovations

As technology continuously advances, the development of big data and statistical methods is also accelerating. The integration of cloud computing, the Internet of Things (IoT), and artificial intelligence technologies makes data collection and analysis more efficient and precise. In the medical field, this means faster disease monitoring and alerting, and more accurate design of personalized treatment plans. Furthermore, with enhanced computing capabilities and innovative algorithms, future statistical methods will be able to handle more complex datasets, providing deeper insights, thus promoting more efficient and personalized medical services.

3. Applications of Big Data Statistical Methods in Disease Prediction and Epidemiology

3.1. Disease Risk Assessment Using Big Data

Big data statistical methods play a crucial role in disease risk assessment. By analyzing vast amounts of data on patients' historical health records, lifestyle habits, genetic information, etc., statistical models can identify high-risk groups and predict the likelihood of specific diseases. For example, using logistic regression analysis, researchers can determine which factors are closely associated with chronic diseases such as heart disease and diabetes, thereby providing a basis for early intervention and preventive treatment.

3.2. Data Mining Techniques in Epidemiological Research

Data mining techniques are increasingly used in epidemiological research, especially in the analysis of outbreak and spread patterns of diseases. By mining and analyzing data from social media, medical devices, and online health portals, researchers can monitor disease transmission trends in real-time, identify hotspots of outbreaks, and evaluate the effectiveness of control measures. Additionally, machine learning techniques like random forests and support vector machines are used to predict disease outbreaks and trends, enhancing the precision and speed of responses [2].

3.3. Case Study: Using Statistical Methods to Track the Spread of Epidemics

During the recent COVID-19 pandemic, statistical methods played a crucial role in analyzing and tracking the virus's transmission pathways. Through integrated models and network analysis, researchers were able to map out the virus's transmission networks, identify super-spreaders, and assess the effectiveness of various public health interventions. For instance, using time-series analysis to predict the peaks and declines of the pandemic, thereby providing a scientific basis for the allocation of medical resources and the formulation of public health policies.

DOI: 10.54254/2755-2721/87/20241546

4. Applications of Big Data in Medical Resource Optimization and Management

4.1. Data-Driven Models for Medical Resource Allocation

In the area of medical resource optimization, big data statistical methods provide an efficient data-driven decision-making model. These models help medical institutions and policymakers optimize resource distribution by analyzing data on medical service usage, patient geographical distribution, and disease prevalence trends. For example, predictive models forecast future patient numbers and medical needs in a specific area, allowing for the proactive deployment of medical personnel and equipment to manage potential medical pressures.

4.2. Optimization of Medical Costs and Service Efficiency

Big data technology also plays a key role in controlling medical costs and enhancing service efficiency. Through statistical analysis, medical institutions can identify cost-driving factors, assess the cost-effectiveness of different treatment options, and optimize service processes. For instance, by analyzing data from patient treatment processes, it is possible to determine which medical procedures are necessary and which may lead to resource wastage, thereby improving the overall efficiency and quality of medical services.

4.3. Case Study: The Effectiveness of Statistical Models in Hospital Management

In a case study at a comprehensive hospital, the introduction of data analysis and statistical models enabled the hospital's management to more accurately predict daily patient flows and optimize manpower resources in emergency and outpatient departments. Moreover, detailed analysis of medical service processes helped the hospital identify efficient and inefficient service elements, significantly improving service quality and reducing operational costs [3].

5. Personalized Medicine and the Integration with Big Data

5.1. From Data Analysis to Personalized Treatment Plans

Personalized medicine is one of the significant applications of big data technology. By analyzing a patient’s genetic information, lifestyle, medical history, and other health-related data, doctors can tailor specific treatment plans for each patient. Statistical methods play a key role in this process by providing precise data analysis, predicting the effectiveness of drugs for specific patients, or identifying potential side effects of treatment methods. This approach not only enhances treatment effectiveness but also significantly reduces medical costs and patient risks.

5.2. Predictive Models Based on Patient Data for Treatment Outcomes

Predictive models developed using statistical methods can accurately forecast treatment outcomes, which is vital for disease management and treatment decisions. For example, by analyzing clinical data and genotypes of cancer patients, predictive models can forecast chemotherapy responsiveness, assisting doctors in choosing the most suitable treatment plan. This method not only enhances the level of treatment personalization but also improves survival rates and quality of life for patients [4].

5.3. Technological Innovations and Future Trends: Intelligent Health Monitoring Systems

With the development of the Internet of Things and artificial intelligence technologies, intelligent health monitoring systems are becoming a significant trend in personalized medicine. These systems continuously monitor a patient’s health status, collecting and analyzing various physiological parameters, such as heart rate and blood pressure. By statistically analyzing this data, the systems can promptly alert potential health issues and automatically adjust treatment plans. This not only enhances the efficiency of medical services but also greatly facilitates patients’ daily lives.[5]

DOI: 10.54254/2755-2721/87/20241546

6. Conclusion

This study, through in-depth exploration of the application of big data and statistical methods in the medical and health field, reveals how these technologies revolutionize modern medical practices, especially in areas like disease prediction, epidemiological research, medical resource optimization, and personalized medical services. Big data not only optimizes the medical decision-making process, enhances the precision and efficiency of treatment plans but also provides robust support for public health management. However, as technology rapidly evolves and its applications expand, ensuring data security and privacy, addressing ethical issues in data analysis, and further enhancing the accessibility and acceptability of analysis techniques remain critical challenges for future research and practice. These challenges also offer new research directions, indicating that the field of medical data science will continue to evolve and bring more innovations.

References

[1] Vahid S , Gholamreza M , Leila O , et al. An MCDM approach to assessing influential factors on healthcare providers' safe performance during the COVID-19 pandemic: Probing into demographic variables [J]. Journal of Safety Science and Resilience, 2023, 4 (3): 274-283.

[2] Faye C , David P , Dorothea N . A systematic review of statistical methodology used to evaluate progression of chronic kidney disease using electronic healthcare records. [J]. PloS one, 2022, 17 (7): e0264167-e0264167.

[3] Health and Medicine - Medical Statistics; Findings from National Cancer Institute (NCI) in the Area of Medical Statistics Described (An Imputation Approach for Fitting Two-part Mixed Effects Models for Longitudinal Semi-continuous Data) [J]. Computer Weekly News, 2020, 364-.

[4] Alex B , Paul A . Statistical Methods for Healthcare Performance Monitoring[M]. Taylor and Francis;CRC Press: 2016-08-05. DOI:10.1201/9781315372778.

[5] John D, Robert S, Emily W, et al. Big Data Analytics in Healthcare: Statistical Methods and Applications [J]. International Journal of Medical Informatics, 2022, 145: 104280. DOI:10. 1016/j.ijmedinf.2021.104280.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241597

An analysis of the hot hand phenomenon in basketball and mid-range shooting

Haoran He

Department of Statistics, University of Pittsburgh, Pittsburgh, 15213, The United States

q56586668@gmail.com

Abstract. Does the Hot Hand phenomenon exist during basketball, especially in the NBA? This question has been controversial over the past years in the sports field. The enormous literature on the Hot Hand effect in basketball was conducted to investigate the Hot Hand hypothesis. This study analyzes this question using a novel data set of all mid-range shots from the 2023-2024 NBA regular season, combined with data on teams and players. A model was built based on the definition of the Hot Hand and potential influential variables such as distance and location to analyze the players' shot data. Supervised machine learning methods fit the model and measure its performance. The results suggest that mid-range shooting streaks do not affect making the next shot in game situations, indicating that the Hot Hand effect is not present in mid-range shooting situations.

Keywords: Hot Hand, Basketball, 2023-2024 NBA regular season, Mid-Range Shot.

1. Introduction

The "hot hand" (also known as the "hot hand phenomenon" or "hot hand fallacy") is a previously considered cognitive social bias that suggests a person who has achieved success in a particular task is more likely to be successful in future attempts. This concept is commonly associated with sports and skill-based activities and originally comes from basketball, where a player is believed to have a higher chance of scoring if they have made successful shots in succession, known as having the "hot hand".

The key point at the center of research into the hot hand is whether the widespread belief that previous shooting streaks increase a player's chance of hitting the next shot is true. If a NBA player makes two, three or four shots in a row, then he is hot handed and is more likely to make his next shot than expected. This study started with Gilovich et al. [1], who demonstrated that the phenomenon was a cognitive illusion caused by random sequences. For example, when a person flipped a coin ten times, there was a chance of having five consecutive heads. However, it was difficult to conclude that the previous four heads increased the chance of getting a head in the fifth attempt. Thus, Gilovich et al. provided that there was no sufficient evidence for a correlation between successive shots and the chance of making the next shot. Then, William O. Brown and Raymond D. Sauer [2] set a point-spread pricing model to prove that belief in the hot hand affected the point-spread betting market, but their study didn't support the hot hand in a real-world context.

Recent research into the hot hand has usually concentrated on controlled settings such as shooting experiments, the NBA 3-point contest, 3-point shooting, and free-throw shooting in games. Both Arkes

DOI: 10.54254/2755-2721/87/20241597

[3], Yaari and Eisenmann [4] found evidence of the hot hand phenomenon in free throws. Miller and Sanjurjo [5-7] demonstrated the hot hand in controlled and semi-controlled settings in their three papers. At the same time, recent experiments have tested for a hot hand in the run of play in NBA games. Bocskocsky et al [8], and Csapo and Raab [9] found that both offences and defences react to made shots. In the run of play, Lantis and Nesson analyzed detailed data on free throws [10] and the NBA 3-point contests [11]. They found a small hot hand effect for free throws and 3-point shots within shot locations, while they had the opposite results for field goal attempts across shot locations. Studies have continued to follow up and revived the hot-hand debate in academia. Kostas Pelechrinis, an associate professor of computing and information at Pitt's School of Computing and Information, claimed that Hot hand exists [12] and "players can indeed get hot in actual live-game situations [13]."

The study aimed to empirically analyze whether the hot hand exists in the NBA for mid-range shots. The author used detailed data for the 2023-2024 NBA regular season. The data from modern-era basketball provides a large sample size and better quality in terms of specificity relative to the datasets used before. Thus, this project expects a better outcome and interpretation of hot hand with a modern dataset of shooting records of renowned NBA players. Moreover, previous papers only considered 3-point shots and free throws in games, but they ignored mid-range shots, which were 2-point shots and more likely to reflect the hot hand effect. Therefore, this paper focused on an analysis of the hot hand effect for the mid-range shots.

2. Data and Methods

2.1. Data

2.1.1. Data Overview

The main data source consists of play-by-play data for the 2023-2024 NBA regular season from the nbastatR package, which is downloaded from Github. This package combines data from credible sources, including NBA's API, HoopsHype, nbadraft.net, and Basketball Reference.com. One of the datasets provided by the package is game_logs, which contains a history of in-game events. For the 2023-2024 regular season, the game_logs data consists of 58 columns and 26,401 rows. This dataset is used to identify the NBA player with the highest-scoring game, as these players are more likely to have hot hands. The top 5 players who scored the most in a single match were then identified. Additionally, the shot data for five players includes 1487, 1436, 1305, 1652, and 851 rows, respectively, with each player having 27 columns of shot-relevant features such as location coordinates, shot types, and whether the player made the shot or not.

This paper also used another data set called team-shots, which provided the teams and players for every game and detailed information for every event in each game, including typeEvent, typeAction, typeShot, zoneBasic, isShotMade, and so on.

When the author set zoneBasic = "Mid-Range", the team-shots data set could build the mid-range shooting data set of each team and each player. Thus, after data wrangling, the mid-range shooting data set was used to test the existence of the hot hand fallacy.

2.1.2. Exploratory Data Analysis

The study utilized histograms to visually represent the distribution of game scores and players. The goal was to identify players with high scores in a single game, as they were more likely to have "hot hands". The histogram showed that the scores from NBA players followed a right-skewed distribution, with only the NBA All-Star roster scoring more than 25 points in the 2023-2024 regular season. Players such as Luka Doncic, Joel Embiid, Shai Gilgeous-Alexander, Kevin Durant, and Devin Booker were identified as the most likely "hot-hand" mid-range shooting players.

Subsequently, shot charts were used to analyze the shot location and shot types for these players. The analysis revealed that Luka Doncic was likely to make his mid-range jump shots from the left side and left center areas, while Kevin Durant attempted a significant volume of successful mid-range shots in

DOI: 10.54254/2755-2721/87/20241597

the right side and right center areas. Joel Embiid tended to make a majority of jump shots from the high post area, while Shai Gilgeous-Alexander favored mid-range attempts from both wings and near free throw line. Devin Booker was a phenomenal mid-range shooter, exhibiting a high shooting percentage across all areas.

The study utilized Coxcomb charts to illustrate the predominant shot types for each player and their respective proportions. The visual representation indicated that Luka Doncic showed a preference for step-back jump shots and a few turnaround fadeaway shots, while Joel Embiid favored jump shots and pull-up jump shots. Shai Gilgeous-Alexander tended to favor pull-up jump shots and step-back jump shots, while Kevin Durant and Devin Booker displayed a notable tendency for pull-up jump shots.

2.2. Methods

2.2.1. Variables and Model Specification

To test the hot hand, this study built a logistic regression probability model that could represent the definition of Hot Hand. Incorporating these factors as additional controls in the regression was crucial for understanding the impact of other variables on shot success. Without these factors, there could be a risk of omitted variable bias. Furthermore, regression analysis allowed for a natural investigation of subsets of shots to uncover whether a "hot hand" effect was concentrated on specific shot locations or distances. Lastly, through regression, this study was able to adopt a flexible approach for measuring multiple streaks of success over the previous mid-range shots.

This model was shown in the following specification:

$$\Pr(\text{isShotMade} = 1) = \text{logit}^{-1}(\alpha + \beta 1\text{lastShot} + \beta 2\text{lastFive} + \gamma 1\text{locationX} + \gamma 2\text{locationY} + \gamma 3\text{distanceShot}) \quad (1)$$

where $\text{logit}(p) = \ln(p/1-p)$, p was the probability that the player made the current shot, α was the intercept, β was slope for variables of interest, γ was the coefficient for the control variables.

Using this model, the author examined the effects of the previous mid-range shot on the probability of making the current mid-range shot.

In relation to the Hot Hand phenomenon, the response variable "isShotMade" was created to denote whether the player successfully made the attempted shot, which was a categorical variable The predictor variables "lastShot" and "lastFive" were established to capture the player's performance, with "lastShot" indicating the player's success in their previous shot and "lastFive" representing the average shot percentage of the last five shots. Other controlled variables considered in this study encompassed location, distance, name zone, and type of action. Ultimately, the selected covariate variables were "locationX" and "locationY," which signified the precise shot location, and "distanceShot," which denoted the distance of a shot made.

2.2.2. Methodological Approach

The statistical analysis was conducted using R version 4.4.0 and RStudio version 2024.04.0-735. The analysis focused on NBA single-game leaders and records for points during the 2023-2024 regular season, as these NBA players were more likely to make consecutive shots. Using R code, the top single-game leaders for points during the 2023-24 season was identified. Subsequently, exploratory data analysis, such as creating a histogram, was used to describe the game scores from the players.

Before testing the hot hand hypothesis, this paper analyzed each player's shot preference, which indicated their preferred shooting locations on the court. Numerous studies have explored the relationship between shot locations and the hot hand phenomenon. According to Lantis et al. [11], "the difference in shot location from the previous shot also grows in magnitude for longer streaks of success." The findings from studying the hot hand and shot preferences could lead to further insights. For example, if a player was found to have a hot hand and is likely to shoot near the post, this could prompt further investigation into other players who tend to shoot from similar locations.

DOI: 10.54254/2755-2721/87/20241597

To understand how consecutive shooting streaks affect the next shot, this paper analyzed the likelihood of making the next shot based on the previous shooting streak. The fluctuation of the line plot didn't prove the existence of the hot hand. However, if the hot hand does exist, these players would experience an increase in their field goal percentage as they continued to make shots successively.

To test the hypothesis of the Hot Hand, the author built a logistic regression model to examine the association between the current shot and the previous shots. Using the summary() function, the author obtained regression coefficients which encompass estimated coefficients, standard error, z-value, and p-value.

This paper tried to predict the categorical response variable, isShotMadeBinary, using supervised learning algorithms.

First, the author divided the dataset into a training set comprising 75% of the original data and a testing set comprising 25% of the original data. The author then applied some classification methods including regularized logistic regression, random forest, and K-nearest neighbor (K-NN) to train the model. After that, this paper performed model tuning to identify the best parameters for each model. The author used 10-fold cross-validation to determine the appropriate hyper-parameters for each model. For regularized logistic regression, the author optimized the lambda parameter, while the author used grid search to find the best number of randomly selected predictors for the random forest algorithm. The author set the mtry hyper-parameter range from 2 to 16 and found the optimal value for each model. The author set ntree to 1,000 as a large number is generally better. Finally, the author compared the performance of each model by calculating the misclassification rate on the 25% testing set, and predicted the outcome for the testing set.

3. Results

This paper began by examining a simple hypothesis of the hot hand: whether a player who made his last shot was more likely than expected to make his next shot. Calculating the probability of making the next shot based on the previous shooting streak. This paper used the mutate function in R to create new variables: nowShot, previousShot, and shotPercentage. The variable nowShot meant the current status of consecutive shots, the variable previousShot meant the previous status of consecutive shots, and the variable shot percentage was the shooting percentage of shots made. Then the author created line charts for different NBA players to represent shooting percentage change over time. The fluctuation of line charts told us that players who made their previous mid-range shot were between 4 and 10 percentage points more likely to make their next mid-range shot. After making three or four consecutive shots, the shooting percentage would be at a stable level, which was between 8 and 10 percentage higher than expected. The results from these line plots did not provide sufficient evidence to conclude that the hot hand effect existed during the game of basketball. However, it gave reasonable and valid evidence to dive into a deeper analysis of the hot hand effect using the 2023-2024 regular season data set.

Next, this paper explored the logistic regression model for the hot hand effect. The results were shown in Table 1.

Table 1. Logistic regression analysis for the Hot Hand effect

Players	Variables	Summary			
		Estimated Coef	Std. Error	z value	Pr(>|z|)
Luka Doncic	(Intercept)	0.898855	2.572995	0.349	0.727
	lastShot	-0.095912	0.877982	-0.109	0.913
	lastFive	-1.045906	1.939354	-0.539	0.590
	locationX	0.001329	0.003828	0.347	0.728
	locationY	0.026732	0.017937	1.490	0.136
	distanceShot	-0.258224	0.271811	-0.950	0.342
Joel Embiid	(Intercept)	1.992510	2.157415	0.924	0.3557

DOI: 10.54254/2755-2721/87/20241597

Table 1. (continued).

	lastShot	0.563959	0.551546	1.023	0.3065
	lastFive	-0.870642	1.264764	-0.688	0.4912
	locationX	-0.010691	0.004732	-2.259	0.0239
	locationY	0.021724	0.012098	1.796	0.0725
	distanceShot	-0.317614	0.202069	-1.572	0.1160
Shai Gilgeous-Alexander	(Intercept)	0.674901	1.946527	0.347	0.729
	lastShot	-0.465617	0.626987	-0.743	0.458
	lastFive	-0.314106	1.694400	-0.185	0.853
	locationX	-0.001947	0.002704	-0.720	0.472
	locationY	-0.003746	0.006890	-0.544	0.587
	distanceShot	0.020757	0.147214	0.141	0.888
Devin Booker	(Intercept)	5.926E-01	1.127E+00	0.526	0.5990
	lastShot	8.236E-01	4.437E-01	1.856	0.0634
	lastFive	-1.113E+00	8.987E-01	-1.238	0.2157
	locationX	-8.534E-04	1.891E-03	-0.451	0.6518
	locationY	9.812E-06	4.546E-03	0.002	0.9983
	distanceShot	-3.711E-02	8.641E-02	-0.429	0.6676
Kevin Durant	(Intercept)	-1.346623	1.047587	-1.285	0.199
	lastShot	0.227615	0.372767	0.611	0.541
	lastFive	-0.218845	0.801211	-0.273	0.785
	locationX	0.001261	0.001578	0.799	0.424
	locationY	-0.006118	0.004118	-1.486	0.137
	distanceShot	0.125982	0.080553	1.564	0.118

In Table 1, the paper shows that Luka Doncic and Shai Gilgeous-Alexander had negative coefficients for lastShot and lastFive. Conversely, Joel Embiid, Kevin Durant, and Devin Booker had a positive coefficient for lastShot and a negative coefficient for lastFive. This means that Luka Doncic and Shai Gilgeous-Alexander were not influenced by the hot hand effect, while Joel Embiid, Kevin Durant, and Devin Booker were more likely to make their next shot after their last shot. However, it was also found that the hot hand effect was not significant since all P-values were greater than 0.05.

The paper used supervised learning methods to train the model. The author measured the model's performance by calculating the misclassification rate of logistic regression on a 25% testing set. The misclassification rates ranged between 27% and 57%, showing that the model's performance varied across different testing sets. This result suggested that the hot hand effect for mid-range shooting might exist for some NBA players in specific games, but it wasn't a consistent or significant factor in the NBA.

4. Conclusion

After analyzing the 2023-2024 NBA regular season data, this paper found that the "Hot Hand" phenomenon could not be a consistent or significant factor in the NBA. The statistical analysis all showed that factors like distance and location didn't play a significant role in predicting and interpreting players' next shot.

However, the concept of the "Hot Hand" is still complex. While some statisticians argue that it's a myth, there are other factors at play, such as muscle memory, psychological elements and defensive intensity, that haven't been fully accounted for in this paper. Further research should be conducted to

DOI: 10.54254/2755-2721/87/20241597

take these additional factors into account and develop more comprehensive methodologies to conceptualize player performance when taking a shot.

References

[1] Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. Cognitive psychology, 17(3), 295-314.
[2] Brown, W. O., & Sauer, R. D. (1993). Does the basketball market believe in the hot hand? Comment. The American Economic Review, 83(5), 1377-1386.
[3] Arkes, J. (2010). Revisiting the hot hand theory with free throw data in a multivariate framework. Journal of Quantitative Analysis in Sports, 6(1).
[4] Yaari, G., & Eisenmann, S. (2011). The hot (invisible?) hand: can time sequence patterns of success/failure in sports be modeled as repeated random independent trials?. PloS one, 6(10), e24532.
[5] Miller, J. B., & Sanjurjo, A. (2014). A cold shower for the hot hand fallacy (No. 518). IGIER working paper.
[6] Miller, J. B. and A. Sanjurjo (2018). Surprised by the gambler's and hot hand fallacies? A truth in the law of small numbers. Econometrica 86(6), 2019–2047.
[7] Miller, J. B., & Sanjurjo, A. (2021). Is it a fallacy to believe in the hot hand in the NBA three-point contest?. European Economic Review, 138, 103771.
[8] Bocskocsky, A., Ezekowitz, J., & Stein, C. (2014). Heat check: New evidence on the hot hand in basketball. Available at SSRN 2481494.
[9] Csapo, P. and M. Raab (2014). "Hand down, man down." analysis of defensive adjustments in response to the hot hand in basketball using novel defense metrics. PloS One 9(12), e114184.
[10] Lantis, R., & Nesson, E. (2021). Hot shots: An analysis of the "hot hand" in NBA field goal and free throw shooting. Journal of Sports Economics, 22(6), 639-677.
[11] Lantis, R., & Nesson, E. (2024). The hot hand in the NBA 3-point contest: The importance of location, location, location. Journal of Sports Economics, 15270025231222630.
[12] Pelechrinis, K., & Winston, W. (2022). The hot hand in the wild. PloS one, 17(1), e0261890.
[13] Pelechrinis, K (2024). How real is the 'hot hand' in basketball? https://fox8.com/news/nexstar-media-wire/the-hot-hand-is-a-real-basketball-phenomenon-but-only-some-players-have-the-ability-to-go-on-these-basket-making-streaks/

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241561

Research on investment project risk prediction and management based on machine learning

Ziwen Diao

Northeastern University College of Professional Studie, Northeastern University, Boston, 02115, USA

13589492785@163.com

Abstract. In the era of digital economy, digital transformation is an inevitable choice in line with current economic development and national policy trends. Enterprises use new generation digital technologies such as big data, blockchain, cloud computing, artificial intelligence, and financial technology to apply these technologies to various production and business activities. The research on project risk management has gradually introduced the method of intelligent decision-making, using technologies such as big data and artificial intelligence to identify and analyze risks. We use learning models and ensemble learning techniques to predict and manage the risks of investment projects. Through long short-term memory networks (LSTMs), we are able to effectively extract spatiotemporal features from historical investment data to predict future risk dynamics. In order to further improve the accuracy of prediction and the robustness of the model, we introduced the Gradient Booster (GBM) ensemble learning method. These integrated technologies not only optimize the overall performance of the model, but also enhance the adaptability and forecast accuracy of various market changes. In the experimental analysis, we compare the performance of multiple models on real-world investment datasets, and the results show that the ensemble learning method has significant advantages in risk prediction accuracy.

Keywords: Investment risk prediction, Risk management, Machine learning, Long short-term memory network, Gradient booster.

1. Introduction

Enterprises are proactively carrying out digital transformation by using next-generation digital technologies such as big data, blockchain, cloud computing, artificial intelligence, and financial technology, and applying these technologies to key activities such as production, R&D, sales, and operations. This process not only realizes the economies of scale, scope and long-tail effects brought about by digital technology, but also effectively reduces the overall cost, optimizes the matching of supply and demand, and improves the equilibrium level of the economy [1]. Digital transformation further strengthens the value creation capabilities of enterprises, stimulates entrepreneurship, brings significant digital dividends to enterprises, and also promotes the sustainable development of enterprises.

With the advancement of digital transformation, enterprises are facing many opportunities and challenges, especially in the new and complex competitive environment, how to adapt project risk management has become a key issue. Digital transformation has changed the way project decisions are made, enabling digital technologies to accomplish tasks that were previously unattainable. This

DOI: 10.54254/2755-2721/87/20241561

transformation has not only reshaped the model of project management, but also promoted the digitalization of project management to gradually become an industry consensus and began to be implemented, bringing new vitality to enterprises [2].

With the gradual popularization of industrial digitalization and digital industrialization, the process, structure and objectives of project risk management need to be reconsidered. Traditional project risk management relies on people, systems, and responsibility traceability, and its marginal utility is gradually weakening. With the deepening of enterprise digital transformation, projects have become more complex, the amount of data has increased, and the intensity and standards of work have also increased significantly [3]. In order to achieve timely early warning of risks and ensure the efficiency and quality of risk control, project risk management is facing new challenges.

Risk management is gradually shifting from "human control" to "intelligent control", and intelligent risk management methods have become the only way for the development of project risk management. In the process of digital transformation, the reproducibility, easy quantification, and easy transmission of data determine its development trend towards precision, comprehensiveness, and sharing [4]. The characteristics of data sharing and co-governance have prompted scholars to introduce intelligent decision-making methods into the research of project risk management, use new technologies such as big data and artificial intelligence to identify and analyze risks, and adopt data-driven management to improve the efficiency and quality of risk management [5]. Fundamentally, the digital transformation of project risk management means moving from a traditional management model to a new data-driven model.

Enterprises must implement strict risk management in projects that invest in the securities market. Project management should conduct a comprehensive risk analysis at the start-up stage, which can help to control project risks to a certain extent. With the digital transformation of enterprises, this process not only gives enterprises new momentum, but also affects the release of internal information and the information asymmetry with market investors [6]. These changes, in turn, affect a company's risk-taking, risk management, risk prevention, and ability to withstand external uncertainties, which may ultimately affect the stability of stock prices. When managing the risks of corporate investment projects, especially those stocks with a high risk of stock price crash, it is important to take appropriate risk mitigation measures. Through this strategy, the occurrence of stock price crashes can be prevented to the greatest extent, thereby mitigating its negative impact on the business [7]. This proactive risk management not only protects the company's assets, but also enhances the company's ability to respond to market fluctuations.

2. Related Work

Initially, In the field of project risk assessment, scholars at home and abroad have adopted a variety of analysis methods, including Delphi method, analytic hierarchy process, genetic algorithm, artificial neural network, fuzzy mathematics and Bayesian network. Specifically, Dai Yuxiu [8] used the Delphi method (DM) to establish a universally applicable project risk identification checklist for public-private partnership projects. She evaluates and ranks these risk sets through fuzzy mathematical methods, analyzes the risk assessment results accordingly, and finally proposes specific risk response recommendations. On the other hand, Zhang Meng [9] applied the analytic hierarchy process (AHP) to evaluate the risk of investment projects, integrating multiple risk factors to determine the relative importance of each risk management measure in investment risk. These studies not only improve the accuracy of risk management, but also provide a scientific basis for the formulation of risk response strategies.

Additionally, risk identification is a crucial first step in project risk management. Only by doing this preliminary work well can the follow-up risk control and response strategies play a more effective role. Zhang Haiyan [10] pointed out that risk management first needs to anticipate the various types of potential risks, and clarify their types and possible consequences. Although project risks are prevalent in all types of projects and have different probabilities of occurrence, the key is to identify risk points as early as possible to reduce the likelihood of risk occurrence and potential losses. Wang Di [11]

DOI: 10.54254/2755-2721/87/20241561

emphasized that in the investment process, predicting and evaluating different risk categories and their occurrence probabilities is key, and finding favorable factors (FF) to reduce these probabilities will help to deal with risks more effectively in the future. These studies have highlighted the importance of proactive and preventive measures in the risk management process.

In the current research literature, the digital degree index of listed companies has not been widely introduced into the risk management of enterprise investment projects. Most studies rely on traditional methods such as expert scoring and analytic hierarchy process. At the same time, there are relatively few studies that use intelligent decision-making tools, such as machine learning methods, to predict crash risk and manage project risk. This shows that there is still a lot of room for development in the field of risk management using advanced technologies such as machine learning, especially in dealing with complex data and predicting future trends.

3. Methodologies

In this section, the process of risk prediction and management of investment projects using long short-term memory network (LSTM) involves mathematical modeling, data preprocessing, model training, risk prediction and risk management strategy formulation.

3.1. Long short-term memory network

The amnesia gate determines which information should be "forgotten" or discarded from the cellular state. By inputting a combination of the previous hidden state h_{t-1} and the current input x_t into an activation function, the forgetting gate outputs a number between 0 and 1, where close to 1 means "keep this information" and close to 0 means "forget this information". Following Equation 1 describes the proposed process.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

The input gate determines the importance of the new information and the extent to which it should be added to the cellular state, which is expressed as Equation 2. This process involves two parts, one is to use the activation function to determine which values should be updated, and the other is to create a new candidate vector that uses the tanh function to help regulate the network.

$$\widetilde{c_t} = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{2}$$

The cell state is maintained over time, which is slightly adjusted at each time step, and some of the old states are discarded according to the output of the Forgotten Gate, while new candidate values are added. Finally, the output gate controls the output from the cellular state to the hidden state, which is expressed as Equation 3. The output is a version of the tanh of the current cell state (providing a numerical value between -1 and 1), which is regulated by the results of the output gate.

$$h_t = o_t * \tanh(c_t) \tag{3}$$

Normalization or normalization, the input time series data is normalized or normalized to improve the efficiency of model training and prevent gradient vanishing problems. Dataset partitioning divides the entire dataset into a training set, a validation set, and a test set to evaluate the model's generalization ability during training.

3.2. Gradient booster

The prediction performance of the model is gradually enhanced by building a series of decision trees, each of which is trying to correct the errors of the previous tree. This approach relies on optimizing a loss function, usually squared error or logistic loss, and is suitable for regression and classification problems. The goal of gradient booster is to find a function $F(x)$ that minimizes the loss function L, which is expressed as Equation 4.

$$L\big(y, F(x)\big) = \sum_{i=1}^{n} L(y_i, F(x_i)) \tag{4}$$

Gradient boosting is done by initializing the model, which is expressed as Equation 5. The first step of the model is initialized by finding the constant γ that minimizes the loss function.

DOI: 10.54254/2755-2721/87/20241561

$$F_0(x) = arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \tag{5}$$

During the training process of the gradient booster, the critical steps include calculating the negative gradient of the residuals of each data point under the current model predictions, which helps to indicate how the model's predictions can improve. Next, we use these computed residuals to fit a new decision tree. Each new tree is added to correct for the prediction error left by the previous tree. Finally, by adding the predictions of this new tree, the entire model is updated to include a learning rate parameter that controls the degree to which each new tree affects the overall model. This process is iterative gradually, each time trying to reduce the overall prediction error until a predetermined number of iterations is reached or the model performance is no longer significantly improved.

4. Experiments

4.1. Experimental Setups

In this study, we designed an experiment to predict and manage corporate crash risk, using a dataset that includes financial metrics, non-financial metrics, macro impact factors, and degree of digitalization. By cleaning, encoding, and normalizing the data, and evaluating their performance through cross-validation methods. After selecting the optimal model, we train the full data and evaluate the performance using metrics including accuracy, and ROC curves. The results of the model are designed to help enterprises take appropriate preventive measures, such as adjusting financial strategies and optimizing operating models, to reduce the risk of crashes, so as to provide scientific data support for enterprise decision-making. The Kern County dataset contains detailed information about the various bonds issued by Kern County, such as the bond amount, issue date, maturity date, interest rate type, and credit rating of the bond. This dataset can be used to analyze and predict the debt repayment capacity and fiscal health of Kern County under different economic conditions, and also provides valuable empirical data for studying local government debt management and bond market dynamics.

Figure 1 illustrates the distribution of multiple financial indicators related to bond issuance. The subgraphs include: Principal Amount, New Money, Refunding Amount, Issue Costs Pct of Principal Amt, and Total Issuance Costs. Each scatter plot shows the corresponding financial indicator as a function of the total issuance cost, while the histogram shows the frequency distribution of the total issuance cost. Such a visual display helps to analyze and understand the potential impact of different financial indicators on the cost of bond issuance, and supports financial analysis and decision-making.

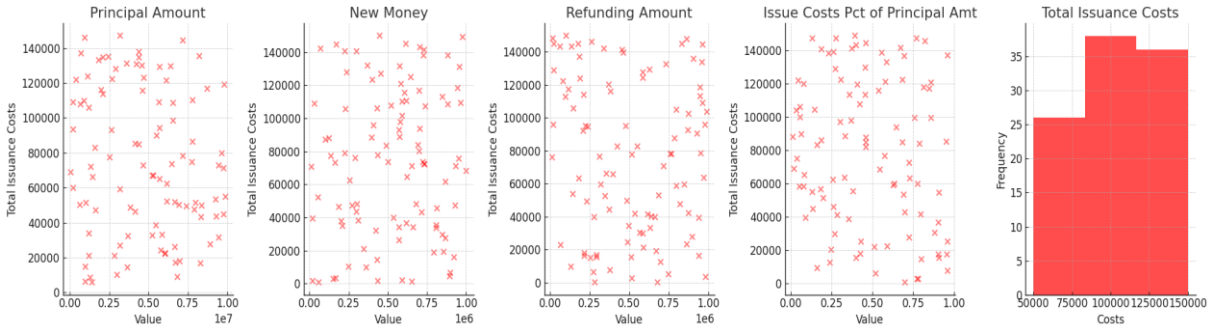

Figure 1. Demonstration of Used Data.

4.2. Experimental Analysis

Accuracy is one of the most basic and intuitive metrics for evaluating the performance of a classification model, and it measures the proportion of all instances of correct classification (true and true negative) to the total sample size. In other words, it reflects how often the model makes the correct judgment across all prediction attempts. Figure 2 shows the risk prediction accuracy comparison results. In a dataset with extremely unbalanced categories, it is possible for a model to achieve a high accuracy rate

DOI: 10.54254/2755-2721/87/20241561

even if it simply predicts all instances as the majority class, but this does not mean that the model has good prediction performance.

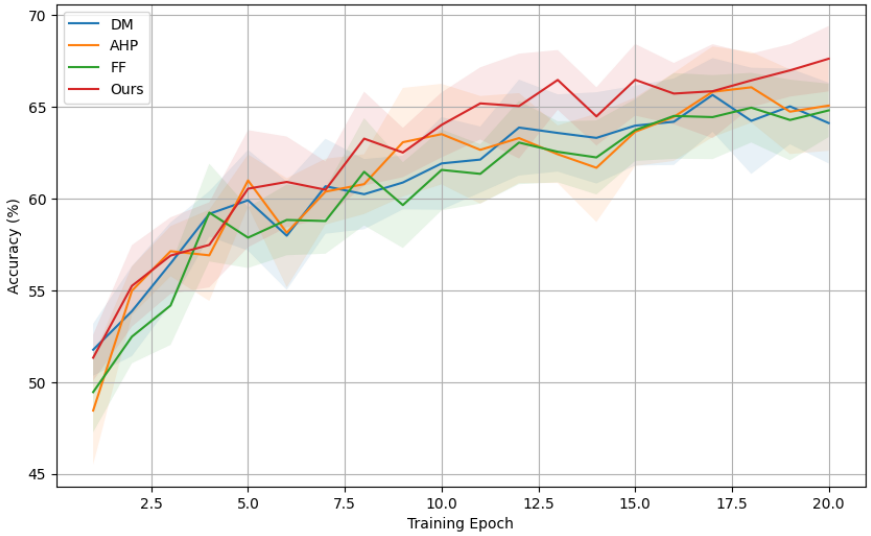

Figure 2. Risk Prediction Accuracy Comparison with Error Shading.

The ROC curve is an evaluation tool for binary classification to evaluate model performance by demonstrating the true rate (TPR) versus false positive rate (FPR) at different thresholds. The true rate reflects the model's ability to recognize positive classes, while the false positive rate reflects how often negative classes are misjudged as positive. Figure 3 shows the ROC curves comparison results.

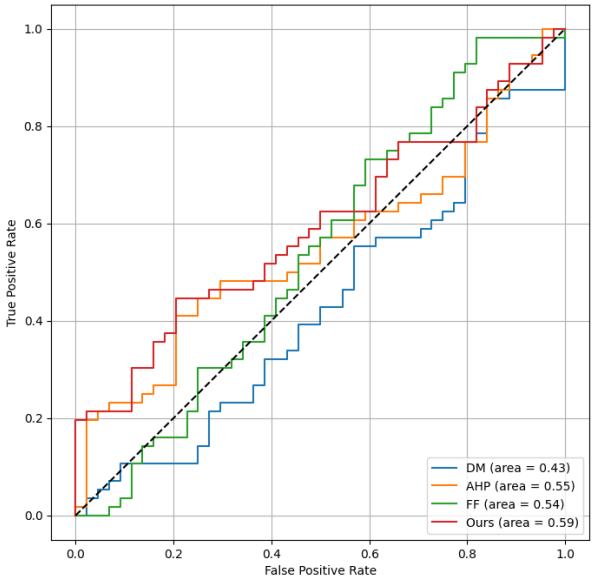

Figure 3. ROC Curve Comparison.

5. Conclusion

In conclusion, by comparing multiple machine learning models including Decision Matrix (DM), Analytic Hierarchy Process (AHP), Fuzzy Frontier (FF), and our proprietary methodology, we found that our method consistently outperforms other models in terms of prediction accuracy and logarithmic loss, showing greater reliability in risk assessment. In addition, the robustness of our method is further emphasized by the application of ROC curves and AUC values, highlighting its effectiveness in

DOI: 10.54254/2755-2721/87/20241561

distinguishing between high- and low-risk items at different threshold settings. In summary, the integration of advanced machine learning technology into investment project risk management can not only make decisions more informed, efficient and effective, but also continue to reveal risk factors in the future, and improve the company's risk mitigation and opportunity capture strategies.

References

[1] Filippetto, Alexsandro Souza, Robson Lima, and Jorge Luis Victória Barbosa. "A risk prediction model for software project management based on similarity analysis of context histories." Information and Software Technology 131 (2021): 106497.

[2] Banerjee Chattapadhyay, Debalina, Jagadeesh Putta, and Rama Mohan Rao P. "Risk identification, assessments, and prediction for mega construction projects: A risk prediction paradigm based on cross analytical-machine learning model." Buildings 11.4 (2021): 172.

[3] Jin, Xin, Qian Liu, and Huizhen Long. "Impact of cost–benefit analysis on financial benefit evaluation of investment projects under back propagation neural network." Journal of Computational and Applied Mathematics 384 (2021): 113172.

[4] Li, Xuetao, Jia Wang, and Chengying Yang. "Risk prediction in financial management of listed companies based on optimized BP neural network under digital economy." Neural Computing and Applications 35.3 (2023): 2045-2058.

[5] Oyedele, Ahmed, et al. "Deep learning and boosted trees for injuries prediction in power infrastructure projects." Applied Soft Computing 110 (2021): 107587.

[6] Zhang, Wen, et al. "Credit risk prediction of SMEs in supply chain finance by fusing demographic and behavioral data." Transportation Research Part E: Logistics and Transportation Review 158 (2022): 102611.

[7] Sun, Xiaolei, Jun Hao, and Jianping Li. "Multi-objective optimization of crude oil-supply portfolio based on interval prediction data." Annals of Operations Research (2022): 1-29.

[8] Dai Yuxiu. Research on whole-process risk management of PPP model engineering projects. MS thesis. Southwest Jiaotong University, (2018).

[9] Zhang Meng. Research on Risk Management of Investment Projects Based on AHP. MS thesis. Zhejiang University, (2015).

[10] Zhang Haiyan. "An Analysis of Financial Risk Management of Venture Capital Companies." Times Finance 1X (2014): 265-265.

[11] Wang Di, and Dai Wei. "Discussion on Corporate Venture Capital Management." Shopping Mall Modernization 4 (2013): 100-101.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241531

Leveraging computational systems for lifecycle management and enhancement of circular economy in fashion: A study on tracking, recycling, and reuse technologies

Yixin Zhou[1], Jiatong Zhao[2,3]

[1]The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Queensland University of Technology, Brisbane, Australia

[3]wellington589125@gmail.com

Abstract. This paper investigates the role of computational systems in managing the lifecycle of fashion products to enhance circular economy practices. By integrating technologies such as IoT, blockchain, and AI, we explore how these systems can track, recycle, and reuse clothing items efficiently. The study highlights the advancements in smart recycling bins, the transparency provided by blockchain, and the optimization of recycling operations through AI. Additionally, the research delves into the impact of online second-hand marketplaces, clothing rental services, and upcycling initiatives, facilitated by computational systems, in promoting sustainable fashion practices. The findings underscore the potential of these technologies to significantly reduce environmental impact, extend garment lifecycles, and foster sustainable consumer behaviors. By leveraging these innovations, the fashion industry can transition towards a more sustainable and circular economy.

Keywords: Circular Economy, Fashion Lifecycle Management, Computational Systems, IoT.

1. Introduction

The fashion industry, known for its rapid production cycles and significant environmental footprint, faces increasing pressure to adopt sustainable practices. A promising solution lies in the concept of a circular economy, which emphasizes the reuse, recycling, and efficient management of resources throughout the lifecycle of products. This study explores how computational systems can facilitate these practices within the fashion sector, offering innovative solutions to long-standing sustainability challenges. Implementing IoT devices for real-time tracking and monitoring of garments can provide valuable data on their condition and usage patterns, enabling timely maintenance and extending their lifespan. Advanced data integration and analysis platforms can aggregate information from various sources, offering comprehensive insights that inform material selection, production techniques, and consumer behavior. These insights can drive more sustainable operations and reduce the environmental impact of fashion products. Furthermore, technologies such as smart recycling bins, powered by IoT and AI, can enhance the efficiency of garment sorting and recycling processes. Blockchain technology can ensure transparency and trust in recycling activities, while AI can optimize sorting and processing operations [1]. Additionally, consumer engagement platforms can educate and motivate individuals to participate in recycling and reuse programs, fostering a culture of sustainability. The development of

DOI: 10.54254/2755-2721/87/20241531

online second-hand marketplaces, clothing rental services, and upcycling initiatives further underscores the potential of computational systems to promote the reuse of fashion products. These platforms not only facilitate transactions and enhance user experiences but also provide valuable data on market trends and consumer preferences, encouraging brands to adopt more sustainable practices. By leveraging these technologies, the fashion industry can make significant strides towards a circular economy, reducing waste, conserving resources, and promoting more responsible consumption and production patterns.

2. Computational Systems in Fashion Lifecycle Management

2.1. Tracking and Monitoring

Implementing computational systems for tracking and monitoring fashion products throughout their lifecycle is a crucial step towards achieving a circular economy. IoT devices, embedded in clothing items, can provide real-time data on the location, condition, and usage patterns of these products. For example, sensors can detect wear and tear, usage frequency, and environmental conditions the garment is exposed to, such as humidity and temperature. This data can be transmitted to centralized systems for analysis, allowing stakeholders to monitor the lifecycle stages and identify opportunities for intervention. For instance, retailers can track the wear and tear of garments to offer timely maintenance services, such as repair or refurbishment options, while consumers can receive notifications for proper garment care tailored to their specific usage patterns [2]. By ensuring continuous monitoring, computational systems can significantly extend the lifespan of fashion products and reduce the frequency of replacements. This proactive approach not only reduces waste but also enhances customer satisfaction by maintaining the quality and longevity of their purchases.

2.2. Data Integration and Analysis

The integration and analysis of data collected from various sources are fundamental for effective lifecycle management. Advanced data analytics platforms can aggregate information from IoT devices, RFID tags, and consumer apps, providing a comprehensive view of each garment's lifecycle. This holistic perspective enables the identification of trends and patterns that can inform decision-making. For example, data analytics can reveal which types of fabrics are most durable or which manufacturing processes result in fewer defects, guiding brands towards more sustainable production choices. Additionally, predictive analytics can forecast future wear and tear, enabling companies to preemptively address potential issues [3]. By leveraging these insights, fashion brands can make informed choices about material selection, production techniques, and maintenance practices, ultimately leading to more sustainable operations and reduced environmental impact. Furthermore, this data can be shared with consumers to promote transparency and trust, showing them the environmental footprint of their clothing choices and encouraging more responsible consumption habits. Table 1 summarizes how data from IoT devices, RFID tags, and consumer apps can be integrated and analyzed to inform sustainable decision-making in fashion lifecycle management, enhancing material selection, production processes, and consumer transparency.

Table 1. Data Integration and Analysis for Sustainable Fashion Lifecycle Management

Data Source	Collected Data	Analytics Outcomes	Impact
IoT Devices	Real-time location, condition, usage patterns	Identify durable fabrics, monitor wear and tear	Inform material selection, enhance garment longevity
RFID Tags	Product identification, lifecycle tracking	Track manufacturing defects, optimize processes	Guide sustainable production, reduce defects
Consumer Apps	User behavior, preferences, feedback	Predict future wear, promote sustainable choices	Encourage responsible consumption, increase transparency

DOI: 10.54254/2755-2721/87/20241531

2.3. Consumer Engagement

Engaging consumers in the lifecycle management of fashion products is essential for fostering sustainable behaviors. Computational systems can facilitate this engagement through personalized apps and platforms that educate and encourage users to participate in recycling and reuse programs. For instance, apps can offer tips on garment care, suggest repair services, and provide information on local recycling centers. These platforms can provide users with information on how to care for their garments, where to donate or recycle them, and the environmental benefits of doing so. Additionally, gamification elements, such as rewards for sustainable actions like recycling or purchasing second-hand items, can motivate consumers to actively contribute to the circular economy. By making sustainability an integral part of the consumer experience, computational systems can drive widespread adoption of eco-friendly practices in the fashion industry. Moreover, social sharing features can enable users to showcase their sustainable actions, further spreading awareness and encouraging community participation in sustainability initiatives [4].

3. Technologies Promoting Fashion Recycling

3.1. IoT and Smart Recycling Bins

The integration of IoT technology with smart recycling bins represents a significant advancement in promoting fashion recycling. These bins, equipped with sensors and connectivity features, can identify and sort clothing items based on material composition and condition. For example, smart bins can use image recognition and RFID scanning to categorize fabrics, distinguishing between cotton, polyester, wool, and other materials. The data collected by these bins can be transmitted to recycling centers, enabling more efficient sorting and processing of garments. This technology not only improves the accuracy of recycling efforts but also reduces the labor and costs associated with manual sorting. Additionally, real-time data from these bins can inform municipalities about the volume and type of recyclables collected, allowing for better planning and resource allocation. By enhancing the efficiency of the recycling process, IoT-enabled smart bins can significantly increase the volume of clothing that is successfully recycled, contributing to the circular economy [5]. Furthermore, these systems can provide feedback to manufacturers and consumers about the recyclability of products, driving improvements in design and consumption patterns. Figure 1 illustrates the efficiency improvement in recycling different types of materials with IoT-enabled smart bins.

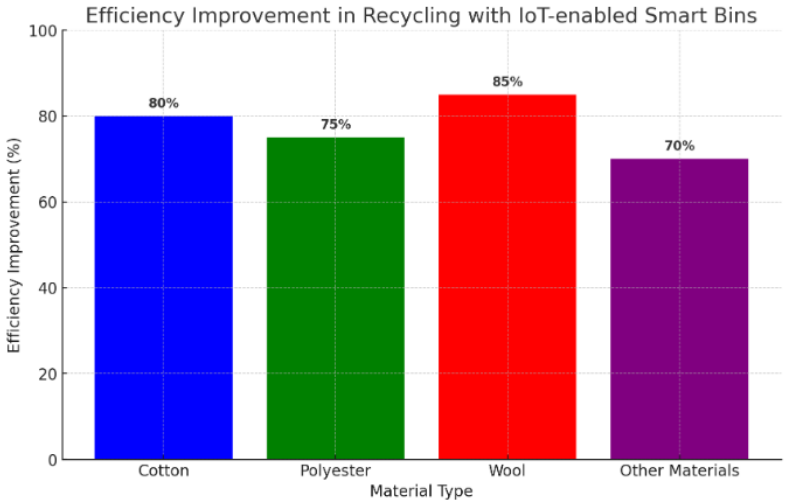

Figure 1. Efficiency Improvement in Recycling with IoT-enabled Smart Bins

DOI: 10.54254/2755-2721/87/20241531

3.2. Blockchain for Transparency and Trust

Blockchain technology offers a robust solution for ensuring transparency and trust in the recycling process. By creating an immutable ledger of transactions, blockchain can track the journey of each garment from donation to recycling to reuse. This transparency allows consumers to verify that their donated items are being recycled responsibly and not ending up in landfills. For example, each step of the recycling process can be recorded on the blockchain, including collection, sorting, processing, and distribution of recycled materials. Additionally, fashion brands can use blockchain to demonstrate their commitment to sustainability by providing verifiable records of their recycling efforts. This increased transparency can enhance consumer trust and encourage more people to participate in recycling programs, thereby supporting the circular economy. Blockchain can also facilitate the development of certification schemes for recycled materials, ensuring that consumers and businesses have access to reliable information about the origins and sustainability of products. By establishing a trusted system for tracking and verifying recycling activities, blockchain can play a crucial role in promoting responsible consumption and production practices in the fashion industry [6].

3.3. AI in Recycling Operations

Artificial Intelligence (AI) plays a pivotal role in optimizing recycling operations within the fashion industry. AI algorithms can analyze data from various sources to improve the efficiency of sorting and processing garments. For example, machine learning models can be trained to recognize different types of fabrics and their recycling requirements, enabling automated sorting systems to operate with greater accuracy. These models can process large volumes of data from sensors, cameras, and other sources to make real-time decisions about how to handle each item. AI can also predict the optimal recycling methods for different materials, minimizing waste and maximizing resource recovery. For instance, AI systems can recommend specific chemical or mechanical recycling processes based on the material composition and condition of garments [7]. By leveraging AI, recycling centers can enhance their operational efficiency, reduce costs, and increase the overall effectiveness of recycling programs. Furthermore, AI can provide insights into the environmental impact of recycling operations, helping to identify areas for improvement and drive the development of more sustainable practices.

4. Enhancing Reuse through Computational Systems

4.1. Second-Hand Marketplaces

The development of online second-hand marketplaces has been significantly boosted by computational systems, facilitating the reuse of fashion products. These platforms use algorithms to match buyers with sellers, optimizing the process of finding and purchasing pre-owned garments. Advanced search and recommendation features, powered by machine learning, help consumers discover items that match their preferences, promoting the reuse of clothing. For example, recommendation systems can analyze user behavior and preferences to suggest items that are likely to be of interest, increasing the likelihood of successful transactions. Additionally, these marketplaces often include features for verifying the authenticity and condition of items, ensuring a trustworthy shopping experience. This can involve user reviews, seller ratings, and even third-party verification services. By making it easier for consumers to buy and sell second-hand fashion, computational systems contribute to reducing waste and extending the lifecycle of garments [8]. Furthermore, these platforms can provide data on market trends and consumer preferences, informing brands about the demand for sustainable and second-hand products, and encouraging them to incorporate more sustainable practices into their business models. Table 2 outlines the key features of online second-hand marketplaces and their impact on promoting the reuse of fashion products [9].

DOI: 10.54254/2755-2721/87/20241531

Table 2. Features and Impact of Second-Hand Marketplaces in Fashion

Feature	Description	Impact
Algorithm Matching	Matches buyers with sellers to optimize the purchasing process.	Increases likelihood of successful transactions.
Advanced Search	Helps consumers discover items that match their preferences.	Promotes reuse of clothing by helping consumers find desired items.
Recommendation Systems	Suggests items based on user behavior and preferences.	Enhances user experience by suggesting relevant items.
Authenticity Verification	Ensures the authenticity of items through various verification methods.	Builds consumer trust in the marketplace.
Condition Verification	Verifies the condition of items for a trustworthy shopping experience.	Ensures quality and condition of pre-owned garments.
User Reviews	Provides reviews from previous buyers for informed decisions.	Informs potential buyers about item quality.
Seller Ratings	Rates sellers to build trust and ensure quality transactions.	Builds a reliable seller base and consumer trust.
Third-party Verification	Offers third-party verification services for additional trust.	Adds an extra layer of trust for buyers.

4.2. Clothing Rental Services

Clothing rental services represent another innovative approach to promoting reuse in the fashion industry. Computational systems play a crucial role in managing the logistics of rental operations, including inventory management, order processing, and delivery scheduling. These systems can track the usage and condition of rented garments, ensuring they are properly maintained and cleaned between rentals. For example, RFID tags can be used to monitor the number of times an item has been rented and its condition after each use, enabling timely maintenance and quality control. Additionally, rental platforms can use data analytics to understand consumer preferences and optimize their inventory accordingly. This can involve analyzing rental patterns to identify popular items and ensure sufficient stock levels, as well as predicting future demand based on trends and user behavior. By providing consumers with access to a wide range of garments without the need for ownership, clothing rental services reduce the demand for new products and support a circular economy [10]. Moreover, rental services can collaborate with designers and manufacturers to create garments specifically designed for durability and multiple uses, further enhancing the sustainability of the fashion industry.

4.3. Upcycling Initiatives

Upcycling, the process of transforming old garments into new products, is greatly enhanced by computational systems. Design software and digital fabrication tools enable designers to create upcycled products more efficiently and with greater precision. For example, CAD (Computer-Aided Design) software can assist in creating detailed designs that maximize the use of existing materials, while minimizing waste. These tools can help identify the best ways to deconstruct and repurpose old garments, minimizing waste and maximizing the value of materials. Additionally, platforms that connect designers with consumers interested in upcycled fashion can facilitate the distribution and sale of upcycled products. These platforms can provide designers with access to a broader market and offer consumers unique, sustainable fashion options. By supporting upcycling initiatives, computational systems contribute to the sustainable reuse of fashion products, reducing the environmental impact of the industry. Furthermore, upcycling can drive innovation in fashion design, encouraging the use of unconventional materials and techniques, and fostering a culture of creativity and sustainability.

DOI: 10.54254/2755-2721/87/20241531

5. Conclusion

In conclusion, the integration of computational systems in the fashion industry presents substantial opportunities for enhancing lifecycle management and promoting a circular economy. Technologies such as IoT, blockchain, and AI can significantly improve the tracking, recycling, and reuse of garments, leading to more sustainable practices. The study has demonstrated the potential benefits and challenges associated with these technologies, highlighting their role in reducing environmental impact and extending the lifecycle of fashion products. Moving forward, continuous innovation and collaboration among industry stakeholders will be crucial in achieving the goals of the circular economy and fostering a more sustainable fashion industry. By embracing these advancements, the fashion sector can contribute to a more sustainable future, reducing its ecological footprint and promoting responsible consumption and production practices.

References

[1] Kirchherr, Julian, et al. "Conceptualizing the circular economy (revisited): an analysis of 221 definitions." Resources, Conservation and Recycling 194 (2023): 107001.

[2] de Oliveira, Carla Tognato, and Giovanna Groff Andrade Oliveira. "What Circular economy indicators really measure? An overview of circular economy principles and sustainable development goals." Resources, Conservation and Recycling 190 (2023): 106850.

[3] Vidal-Ayuso, Fatima, Anna Akhmedova, and Carmen Jaca. "The circular economy and consumer behaviour: Literature review and research directions." Journal of Cleaner Production (2023): 137824.

[4] Piscicelli, Laura. "The sustainability impact of a digital circular economy." Current Opinion in Environmental Sustainability 61 (2023): 101251.

[5] Alberich, Josep Pinyol, Mario Pansera, and Sarah Hartley. "Understanding the EU's circular economy policies through futures of circularity." Journal of Cleaner Production 385 (2023): 135723.

[6] Daukantienė, Virginija. "Analysis of the sustainability aspects of fashion: a literature review." Textile Research Journal 93.3-4 (2023): 991-1002.

[7] D'Itria, Erminia, and Reet Aus. "Circular fashion: evolving practices in a changing industry." Sustainability: Science, Practice and Policy 19.1 (2023): 2220592.

[8] Amasawa, Eri, et al. "Can rental platforms contribute to more sustainable fashion consumption? Evidence from a mixed-method study." Cleaner and Responsible Consumption 8 (2023): 100103.

[9] Malinverno, Nadia, et al. "Identifying the needs for a circular workwear textile management–A material flow analysis of workwear textile waste within Swiss Companies." Resources, Conservation and Recycling 189 (2023): 106728.

[10] Kumar, Sunil, et al. "An optimized intelligent computational security model for interconnected blockchain-IoT system & cities." Ad Hoc Networks 151 (2023): 103299.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241582

The practical applications of federated learning across various domains

Hanjing Wang

College of Information Science and Engineering, Ocean University of China- Ocean University of China, Qingdao, China

whj4421@stu.ouc.edu.cn

Abstract. With the advancement of artificial intelligence technology, a vast amount of data is transmitted during the model training process, significantly increasing the risk of data leakage. In an era where data privacy is highly valued, protecting data from leakage has become an urgent issue. Federated Learning (FL) has thus been proposed and applied across various fields. This paper presents the applications of FL in five key areas: healthcare, urban transportation, computer vision, Industrial Internet of Things (IIoT), and 5G networks. This paper discusses the feasibility of implementing FL for privacy protection in the aforementioned five real-world application scenarios and analyzes its accuracy and effiency. Additionally, it compares the FL framework with traditional frameworks, exploring the improvements FL has made in terms of privacy protection and performance, as well as the existing shortcomings of the FL framework. Further discussions are provided on potential future improvements. Moreover, this paper offers an outlook on current research trends and the developmental prospects in this research field.

Keywords: Federated learning, privacy-preserving, efficiency.

1. Introduction

Privacy issues are one of the major concerns today. The development of big data, artificial intelligence, and other technologies has inevitably led to problems related to data privacy breaches. User data is transmitted during the use of various software, websites, etc.; if this information is leaked, it can lead to various illegal activities such as fraud and extortion. The ability of enterprises to protect user privacy data significantly affects users' trust in them. With technological advancement, an increasing number of devices are being utilized. Currently, nearly 7 billion connected devices are used in Internet of Things (IoT) [1], and the number of smartphone users has almost reached 3 billion [1]. Consequently, the volume of data transmission between devices has greatly increased. In the field of deep learning, a substantial amount of data is collected and utilized for training deep models. While computational power and time have garnered widespread attention, data privacy issues were initially overlooked. The increase in data transmission volume can easily lead to serious privacy breaches [2]. In certain fields, the transmitted data often involves industry secrets and user privacy; if such data is leaked, it could lead to severe incidents, thus drawing significant attention to data security issues.

To safeguard data privacy, researchers have undertaken various attempts, among which the introduction of the Federated Learning (FL) model stands out as a significant approach. Researchers are not only delving deeply into FL model algorithms but also exploring potential real-world applications

DOI: 10.54254/2755-2721/87/20241582

of FL across multiple domains. A critical question they address is how to utilize FL models to protect data privacy without compromising model accuracy. This paper aims to review the applications of FL in the fields of medicine, urban transportation, visual systems, Industrial Internet of Things (IIoT), and 5G networks. It analyzes and summarizes the performance of FL in data privacy protection. Additionally, the paper organizes and examines methods to enhance FL's performance in terms of accuracy and data processing efficiency. Furthermore, it provides an outlook on the future prospects of FL applications.

2. The Theory of FL

To overcome the limitations posed by data privacy on artificial intelligence, FL was proposed to safeguard user privacy [3]. Research on FL is still in its infancy, and many scholars are conducting studies in this field. Overall, FL is an iterative process [4], in which data are brought into the code. FL is implemented through three steps: (1) initiate a global model [4]; (2) training the initial machine learning (ML) models on clients using personal data [4]; (3) training the local models at the client level, updating them, and sending the updates to the server where they are aggregated and used in order to update the global model [4]. After that, the newly updated model is transmitted back to each client [4]. Steps (2) and (3) are repeated [4]. Based on the aforementioned steps, there is no data transmission between clients. The data of each client is kept locally, which helps to protect the privacy of user data. The above steps are depicted in Figure 1.

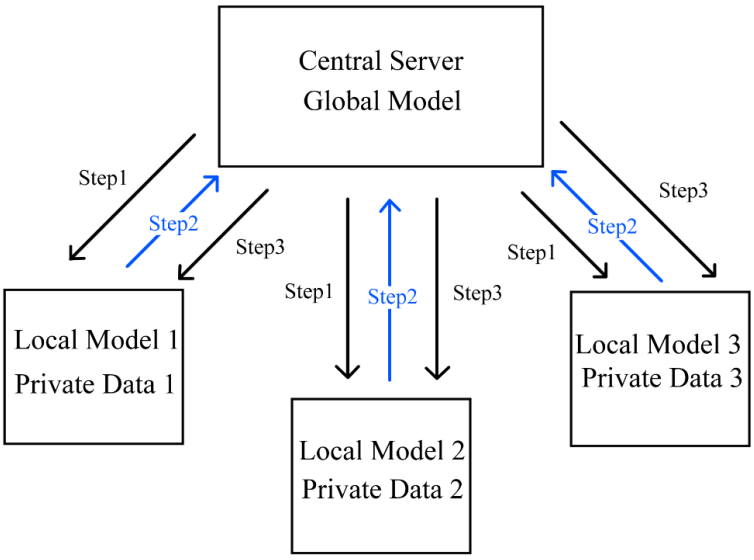

Figure 1. The three steps of FL(Picture credit : Original)

FL can be categorized in various ways, including network topology, data partitioning, open-source frameworks, data availability, and optimal aggregation algorithms [4]. Based on network topology, common classifications of FL are Centralized & Clustered FL and Fully-Decentralized FL [4]. Centralized & Clustered FL relies on a single central server, while Fully-Decentralized FL uses multiple coordinating nodes and clusters for distributed aggregation. According to data partitioning, it can be split into the following three types: Vertical FL, Horizontal FL, and Transfer FL [4]. For instance, SecureBoost, which combines XGBoost and FL, falls under this category. Figure 2 illustrates the specific categorization methods of FL [4].

DOI: 10.54254/2755-2721/87/20241582

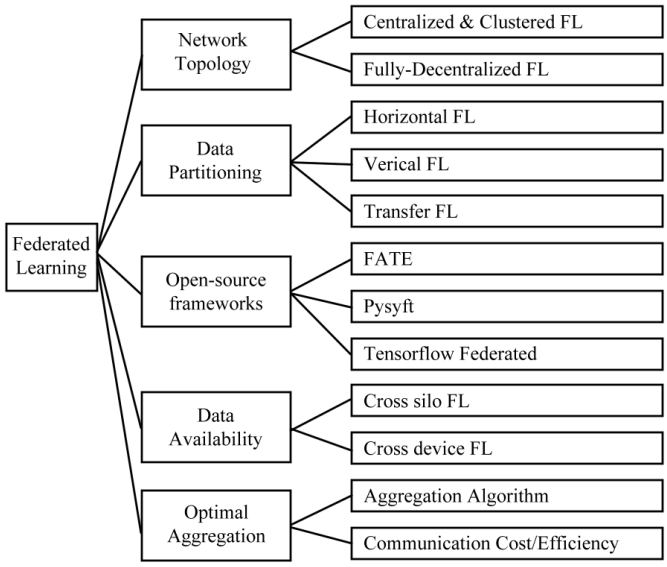

Figure 2. The Classification of FL[4]

FL data often exhibits authenticity and privacy, with labels that can be directly obtained [3]. In the real world, most domains involving private data receive considerable attention, and many scenarios involve high data transmission costs or require distributed collaborative training (e.g., intelligent transportation and autonomous driving). Data in the fields of medicine, urban transportation, vision, IIoT, and 5G networks often align with the characteristics required for FL data. Numerous scholars have conducted application research on FL in these domains.

3. Application Analysis

In various fields, data privacy and security hold paramount importance. This paper will discuss and summarize the applications of FL in five key areas: healthcare, urban transportation, computer vision, IIoT, and 5G networks, and analyze its performance in these domains.

3.1. Medicine

In the medical field, data often possesses a high degree of privacy and sensitivity [5]. This data includes patients' personal identification information and health information, making the confidentiality of medical data particularly crucial.

To protect user privacy, reference [5] utilized the characteristic of FL where data does not need to be uploaded to a central server to train the learning model, while reference [6] employed FL to train local datasets from different sites. The datasets used in references [7, 8] consisted of 120 samples with six different attributes (urinary pain, urethral burning sensation, itching and swelling, urgency, occurrence of nausea, discomfort in the lower back, and body temperature). Reference [6] compared traditional ML methods with FL g methods. To ensure model accuracy while safeguarding data privacy, researchers used datasets in text file format and preprocessed the data. Experimental data from reference [6] indicates that compared to traditional ML approaches, FL not only enhances data privacy but also achieves nearly 100% accuracy.

The researchers in reference [9] utilized a multimodal FL model to evaluate the diagnostic efficacy in gynecologic malignancies, encompassing a dataset of over 500 patients. Figure3 illustrates the process of multimodal FL [10]. In addition to employing the FL model, reference [9] implemented stringent access controls, restricting access to authorized researchers and medical personnel only, who could access patient privacy data under authorization, and anonymized patient data (such as removing identifiable information that could disclose patient identities). Through these methods, the researchers enhanced data privacy protection, offering a feasible approach. Reference [9] partitioned the dataset into

DOI: 10.54254/2755-2721/87/20241582

two parts (training data sets and testing data sets) to leverage the advantages of FL, demonstrating that FL's capability to effectively safeguard data privacy played a crucial role. Table 1 illustrates the performance comparison between traditional methods and multimodal FL [9]. The comparison reveals that multimodal FL enhances sensitivity while safeguarding patient privacy, underscoring its significant potential as a more effective diagnostic tool in the field of gynecologic malignancies.

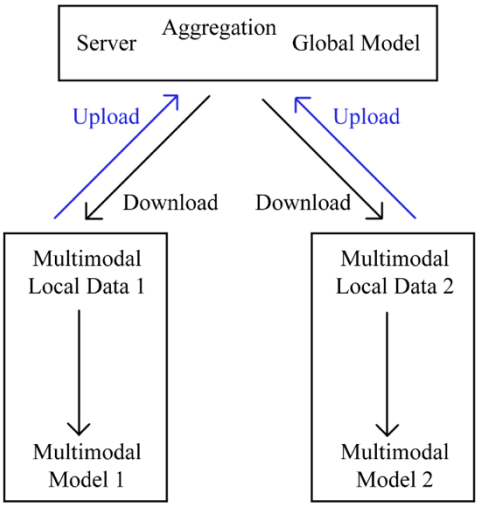

Figure 3. The process of Multimodal FL[10]

Table 1. Traditional method v/s Multimodal FL

	Traditional (CT)	Traditional (MRI)	Multimodal FL Without Image Information	Multimodal FL With Image Information (MRI)
Sensitivity	0.3263	0.359	0.923	0.941
Specificity	0.9215	0.9337	0.922	0.967

3.2. Urban Transportation

In the domain of urban transportation, directly collecting user information may expose their privacy, such as personal identification details, geographical locations, and mobility patterns. The protection of data privacy is directly correlated with the trustworthiness of users.

To guarantee individual data security, reference [11] proposed the DRLE framework to establish a decentralized learning edge computing approach, yet there still exists certain risks during the collection of raw vehicle data [12]. In order to augment the security of data privacy, reference [12] considered employing FL for model training. Meanwhile, to analyze the feasibility of this approach, researchers utilized the same model as reference [13] as a benchmark for comparison. The final results obtained are depicted in Table 2 [12,13]. From the comparison, it is evident that while the privacy of the data has been improved, the accuracy of the proposed model decreased from 76.81% to 71.02%.

Table 2. Baseline v/s FL

	Baseline	FL
Accuracy	0.7681	0.7102
Precesion	0.7681	0.7002
Recall	0.7681	0.7085
F1-Score	0.768	0.7100

DOI: 10.54254/2755-2721/87/20241582

FL's integration and Transport Mode Inference (TMI) was proposed by reference [14] to enhance data privacy, termed as PPDF-FedTMI. To assess the model's performance, researchers utilized a GPS-based dataset [15] and reconstructed its trajectories in experimental preparation. Analysis of the experimental results [14] regarding metrics demonstrating significant potential. However, there is still a need for further development in balancing privacy and utility aspects [14].

3.3. Visual System

In the field of computer vision, data privacy is of paramount importance. For instance, data utilized in applications such as facial recognition and surveillance systems often involve sensitive information like users' personal identities. If this information is compromised, it could potentially lead to malicious events such as identity theft, resulting in economic and psychological losses for the users.

Reference [16] addresses the unique needs of individuals with hearing impairments by leveraging FL to detect Bengali Sign Language while preserving user privacy. Researchers in [16] established a FL framework and conducted a comprehensive evaluation of six models concerning accuracy, precision, F1 score, recall, and loss. Among these, the proposed Federated Averaging (FedAVG) model achieved an accuracy of 98.36% (correctly predicting 9246 out of 9400 samples) [16]. The implementation process of FedAVG is illustrated in Figure 4 [17]. In this experiment, FL demonstrated its capability to effectively protect data privacy while achieving high accuracy. However, there are still some privacy risks associated with collaborative model training that need to be addressed (such as The risks associated with collaborative training participants [16]).

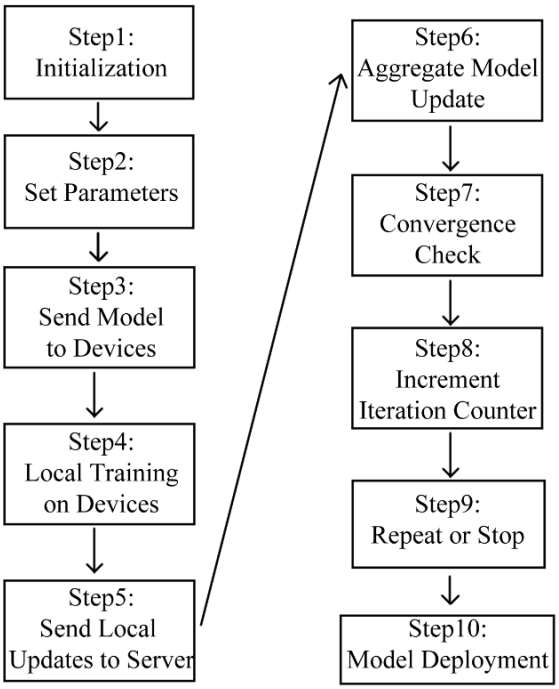

Figure 4. The implementation process of FedAVG [17]

In addition to gesture recognition, FL has also shown promise in human body posture recognition. Reference [18] proposed a FL framework (FL-HPR) for human posture recognition, aiming to protect data privacy. In the study conducted by reference [18], the researchers performed five-fold cross-validation on the client-side performance of three FL models (FedAVG, Fedprox, and FedBN) and found that the FL framework successfully optimized point cloud segmentation networks' average performance while protecting data privacy. In this study, human posture images were either unobstructed or

DOI: 10.54254/2755-2721/87/20241582

minimally obstructed. The researchers anticipate achieving high accuracy in recognition even under conditions of severe occlusion .

3.4. IIoT

In the IIoT domain, data privacy is equally paramount. Data encompassing sensitive business information such as operational efficiency, equipment parameters, and manufacturing processes is involved. Unauthorized access to this data could inflict significant economic losses on enterprises, and in malicious attack scenarios, compromise production processes leading to security incidents.

To mitigate potential data security issues, researchers consider adopting a decentralized architecture, namely the FL model. The study by researchers in reference [19] integrates FL methods to achieve large-scale distributed deep learning in IoT environments, ensuring user privacy protection and efficient communication. Reference [19] primarily employs approximate computing, distributed optimization algorithms, incremental learning, and differential privacy techniques, among others, to test three real-world datasets. Ultimately, reference [19] achieves a privacy preservation accuracy of 98%, marking an improvement over traditional privacy protection techniques, while also enhancing communication efficiency. However, its resilience against potential interference attacks requires further enhancement.

Regarding the aforementioned issues, reference [20] conducted a more in-depth investigation. To safeguard data privacy against threats such as data poisoning attacks and interference attacks, reference [20] researchers proposed a FL model using multiparty computation. FL faces various threats including interference attacks, data leakage, and model reverse engineering, as illustrated in Figure 5 [20]. Given these myriad threats, upgrading FL systems becomes imperative. Reference [20] employed algorithms based on secure multiparty computation to facilitate collaborative computation among multiple parties while ensuring their respective data's privacy. Compared to conventional FL algorithms, it achieved enhanced accuracy; however, an increase in the number of clients also led to higher communication overhead and latency rates compared to traditional FL algorithms.

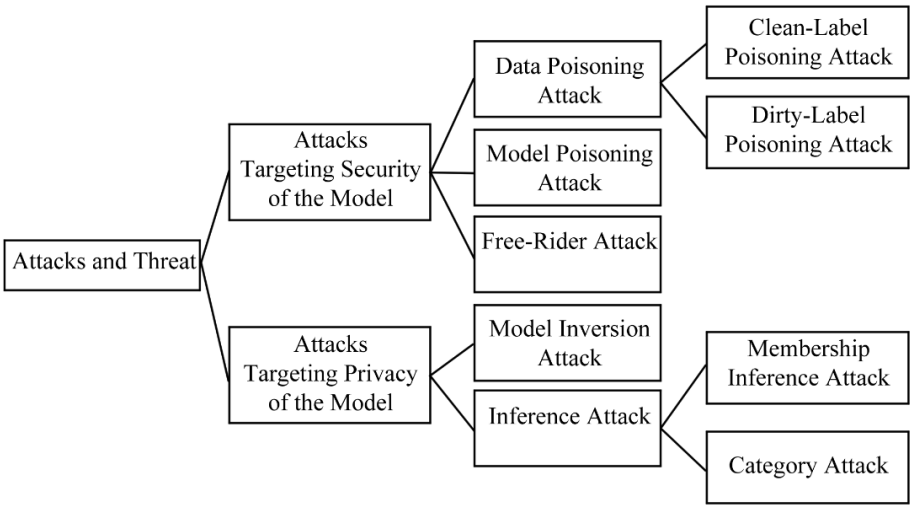

Figure 5. The classification of threats faced by FL[20]

3.5. G Networks

With the widespread adoption of 5G networks, a vast amount of information concerning user privacy and corporate confidentiality is transmitted through these networks. While 5G networks facilitate rapid data processing, they also escalate the risk of malicious attacks. Therefore, safeguarding the data transmitted over 5G networks is of paramount importance.

The researchers in reference [17] addressed privacy protection concerns by integrating FedAvg, adaptive learning rate, and secure aggregation for collaborative model training. Compared to traditional methods such as decision trees and linear regression, the approach proposed in reference [17]

DOI: 10.54254/2755-2721/87/20241582

demonstrates significantly superior accuracy (achieving 95.2%) while also preserving data privacy. Furthermore, it exhibits higher efficiency in data processing, meeting real-time requirements and utilizing memory resources more effectively than methods like logistic regression [17]. Table 3 illustrates the comparative performance of FL versus traditional methods [17]. In the realm of network applications, FL models exhibit exceptional adaptability and hold promising prospects for broader future applications.

Table 3. Proposed FL Method v/s Traditional Method

Method	Accuracy (%)	Efficiency (ms)	Memory Usage (MB)	Scalability (Nodes)
Proposed FL Method	95.2	25	120	5000
Linear Regression	88.5	30	----	----
Decision Trees	92.1	35	----	----
Logistic Regression	----	----	150	4000
Random Forest	----	----	200	3500

Introducing artificial intelligence algorithms in 5G networks raises significant concerns about privacy protection. To effectively safeguard user data privacy, researchers in [21] proposed an asynchronous weight updating framework based on FL. This framework is split into two distinct parts: client-side training and central node training, allowing clients to locally update their model parameters and optimizing them on network slicing [21]. As the number of clients increases, reference [21] demonstrated stable performance improvements, with the model's performance also increasing with additional time rounds. Reference [21] achieved a low-latency approach that reduces overhead while enhancing throughput, demonstrating high-performance applications in the 5G domain.

4. Conclusion

As a newly emerged technology, FL has successfully contributed to privacy protection. This paper provided a brief introduction to the principles of FL, shifting their focus from theoretical aspects to practical applications. The paper organized, analyzed, and summarized the applications of FL in five domains: medicine, urban transportation, visual systems, IIoT, and 5G networks. The findings revealed that FL can effectively protect data privacy, demonstrating superior performance in this regard. Accuracy, a crucial metric for FL models, can be ensured through the integration of other algorithms, model optimization, and data preprocessing. Compared to traditional deep learning models, FL models can achieve significant improvements in accuracy. Additionally, performance metrics such as memory consumption can be enhanced by optimizing certain aspects of FL, such as network slicing. In practical applications, optimized FL models outperform traditional deep learning models concerning privacy protection, data processing efficiency, and accuracy. The optimization direction of FL varies depending on the specific domain and requires contextual judgment. Overall, FL has broad application prospects in real life. Beyond specific domains, researchers can delve deeper into hybrid fields for further exploration.

References

[1] Lim, W. Y. B., et al. 2020 "Federated Learning in Mobile Edge Networks: A Comprehensive Survey," IEEE Commun. Surv. Tutorials, vol. 22, no. 3, pp. 2031-2063.

[2] Fang, C., et al. 2022 “A privacy-preserving and verifiable federated learning method based on blockchain”, Comput. Commun., vol. 186, pp. 1-11.

[3] McMahan, H. B., et al. 2016 “Communication-Efficient Learning of Deep Networks from Decentralized Data”. arXiv preprint arXiv:1602.05629.

[4] Mothukuri, V., et al. 2021 “A survey on security and privacy of federated learning”, Future Gener. Comput. Syst., vol. 115, pp. 619-640.

[5] Shah, U., et al. 2021 "Maintaining Privacy in Medical Imaging with Federated Learning, Deep Learning, Differential Privacy, and Encrypted Computation," 2021 6th Int. Conf. for Convergence in Technol. (I2CT), Maharashtra, India, pp. 1-6.
[6] Shah, H., Patel, R., and Tawde, P. 2023 "Federated Learning to Preserve the Privacy of User Data," 2023 Somaiya Int. Conf. on Technol. and Inf. Manag. (SICTIM), Mumbai, India, pp. 23-27.
[7] Czerniak, J. and Zarzycki, H. 2003 "Application of rough sets in the presumptive diagnosis of urinary system diseases", Artif. Intell. and Security in Comput. Syst., ACS'2002 9th Int. Conf. Proc., Kluwer Acad. Publ., pp. 41-51.
[8] UCI Machine Learning Repository: –Acute Inflammations Data Set.
[9] Hu, Z., et al. 2023 "Comparison of Multi-Modal Federated Learning Framework and SPSS in the Evaluation of Lymph Node Metastasis Probability in Gynecological Malignancies," 2023 IEEE 4th Int. Conf. on Pattern Recognit. and Mach. Learn. (PRML), Urumqi, China, pp. 280-284.
[10] Lin, Yi-Ming, et al. 2023 Federated Learning on Multimodal Data: A Comprehensive Survey, MIR, 20(4): 539-553.
[11] Zhou, P., et al. 2021 DRLE: Decentralized reinforcement learning at the edge for traffic light control in the IoV, IEEE Trans. on Intell. Transp. Syst., vol. 22, pp. 2262–2273.
[12] Gomes, G. L., da Cunha, F. D., and Villas, L. A. 2023 "Differential Privacy: Exploring Federated Learning Privacy Issue to Improve Mobility Quality," IEEE Latin-Am. Conf. on Commun. (LATINCOM), Panama City, Panama, pp. 1-6.
[13] Wang, S. 2018 Traffic jam prediction hackntx. Accessed: 2023-09-29.
[14] Huang, Qihan, et al. 2023 PPDF-FedTMI, A Federated Learning-based Transport Mode Inference Model with Privacy-Preserving Data Fusion, Simul. Model. Pract. and Theory, vol. 129, Art. 102845.
[15] Zheng, Y.,W. Y. B. Lim et al., "Federated Learning in Mobile Edge Networks: A Comprehensive Survey," in IEEE *Commun. Surv.* Tutorials, vol. 22, no. 3, pp. 2031-2063, thirdquarter 2020
[16] Sarkar Diba, Bidita, et al. 2024 Explainable federated learning for privacy-preserving bangla sign language detection, Engineering Applications of Artificial Intelligence, vol. 134, p. 108657.
[17] Ojha, A. C., Yadav, D. Kumar, and B, A. 2023 "Federated Learning Paradigms in Network Security for Distributed Systems," 2023 IEEE Int. Conf. on ICT in Bus. Ind. & Gov. (ICTBIG), Indore, India, pp. 1-5.
[18] Wang, Jiaxin, et al. (2024) Multi-sensor fusion federated learning method of human posture recognition for dual-arm nursing robots, Inf. Fusion, vol. 107, Art. No. 102320.
[19] Du, W., et al. "Approximate to Be Great: Communication Efficient and Privacy-Preserving Large-Scale Distributed Deep Learning in Internet of Things," IEEE Internet Things J., vol. 7, no. 12.
[20] Huang, R.-Y., Samaraweera, D., and Chang, J. M. 2023 "Exploring Threats, Defenses, and Privacy-Preserving Techniques in Federated Learning: A Survey," Computer, vol. 57, no. 4, pp. 46-56.
[21] Bedda, K., Fadlullah, Z. M., and Fouda, M. M. 2022 "Efficient Wireless Network Slicing in 5G Networks: An Asynchronous Federated Learning Approach," 2022 IEEE Int. Conf. on Internet of Things and Intelligence Systems (IoTaIS), BALI, Indonesia, pp. 285-289.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241567

Application of deep learning models in the identification and screening of fake news

Hongchen Zhu

Big Data, Singapore Institute of Management Singapore, Singapore

hzhu014@mymail.sim.edu.sg

Abstract. As data sets and data streams continue to expand, traditional machine learning is becoming less effective in predicting fake news. This paper is a review of deep learning in fake news detection and prevention. Author takes the model based on convolutional neural network as an example to illustrate the principle and application of deep learning in fake news detection, including OPCNN-FAKE, Dual-channel Convolutional Neural Networks with Attention-pooling (DC-CNN) model which is completely based on Convolutional Neural Network (CNN), and Convolutional Neural Network-Long Short Term Memory (CNN-LSTM) model which combines convolutional neural network with long-short time model. These models have obvious advantages in accuracy over traditional machine learning models. This paper then points out the problems of deep learning in the field of fake news identification: it does not have good scalability and slow training speed. The author proposes possible solutions, and widely uses transfer learning and uses distributed computing platforms, such as spark, to train models. Hope this review can help the research on fake news prediction using deep learning.

Keywords: deep learning, neural network, fake news.

1. Introduction

Fake news is a subset of information which are deliberately and strategically constructed lies that are presented as news articles and are intended to mislead the public [1]. An early enough example of how fake news or misinformation to influence human communication is Octavian's propaganda campaign against Antony in the Roman era, which aimed to discredit him by smearing his private life [2]. This kind of false propaganda with strong political overtones is common. With the development of the Internet, the carriers of fake news are no longer limited to books and newspapers, social media has also become one of the main channels for spreading fake news. According to a survey by Andrea Moscadelli, during the covid-19 epidemic, there were more than two million clicks on links to fake news on Italian social media, accounting for 23.1% of the total number of clicks on links related to covid-19 [3]. This range of fake news undoubtedly reduces the efficiency of people obtaining effective information and increases the time cost of obtaining real information. Hence, it is important to identify which information is fake news and remind the public to be vigilant against it.

The development of technology has not only facilitated the spread of fake news, but also helped in combating fake news. Today, researchers are using various data mining methods to improve the model's ability to identify fake. For instance, Shu et al divides the feature extraction of fake news into two categories, 'news content feature' and 'social context features' [4]. Modeling based on news content can

DOI: 10.54254/2755-2721/87/20241567

be categorized into two types: fact-based modeling and style-based modeling, which focuses on the author's style. Additionally, modeling based on social context can be segmented into modeling that considers the author's position and modeling that assesses the dissemination potential of the article. Using these four classification methods, the classification algorithm for fake news is constantly improved. As an emerging technology, deep learning has been proven to be more effective than traditional machine learning in predicting fake news [5] due to its superior feature extraction capabilities in processing high-dimensional and complex data. These capabilities are exactly what is needed for fake news prediction.

Based on this reality, the four classification methods mentioned above have made new technological breakthroughs after the introduction of deep learning. In terms of modeling based on factual information, Xu et al. used graph neural networks to model long-distance semantic relationships in news and evidence [6]. In terms of modeling based on pattern information, Altheneyan et al. created a stacked ensemble model to detect the stance of article titles [7]. At the same time, Sheng et al. proposed the Pref-FEND model, attempting to achieve joint detection of "pattern information-based" and "factual information-based" through graph convolutional neural networks and attention modules [8].

Given the recent important progresses made in this field, this paper aims to summarize the recent deep learning techniques for fake news detection. In the rest of this paper, Section 2 will list the deep learning models and methods used by different researchers. Section 3 will compare and analyze the advantages and disadvantages of different models, as well as explain potential future research directions. Finally, Section 4 will summarize the paper.

2. Method

2.1. Introduction of deep learning

Deep learning is a type of machine learning. Unlike traditional machine learning, deep learning does not require different vector extractors for different tasks, but instead combines simple but nonlinear modules, each of which transforms one level of representation into a higher, more abstract representation [9]. Taking a traditional Artificial Neural Network (ANN) shown in Figure 1 as an example, after the features are passed into the input layer, they are processed by multiple hidden layer functions, and the final result is calculated by the hidden function of the data layer to obtain the final result.

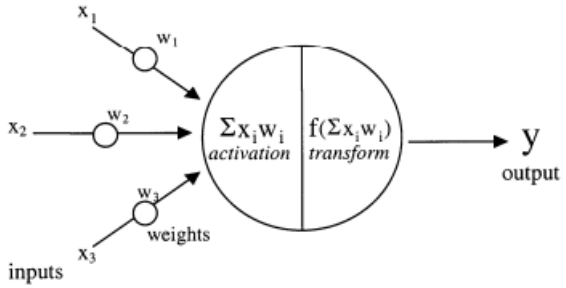

Figure 1. The model of the neural network [10].

After continuous improvement and optimization, deep learning has evolved into different branches to meet different needs. It can be roughly divided into ANN, Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), Graph Neural Networks (GNN), and Recurrent Neural Networks (RNN) according to their structure and application. In natural language recognition, the Long Short-term Memory (LSTM) network improved from the recurrent neural network is a commonly used model. LSTM can capture long-term dependencies by introducing forget gates, input gates, and output gates.

DOI: 10.54254/2755-2721/87/20241567

2.2. OPCNN-FAKE

OPCNN-FAKE is a specialized CNN model for fake news prediction [11]. It extracts high-level and low-level features from news texts through a combination of multiple convolutional layers and pooling layers. The model consists of six main layers. The embedding layer embeds each word of the news text into a vector space, where each row of the vector corresponds to a word. The input dimension represents the size of the vocabulary, and the output dimension represents the dimension of the word vector; the dropout layer is used for regularization to prevent the model from overfitting, and the model dropout ratio is set to 0.5 for the best; the convolution layer is used for feature extraction, and the Rectified Linear Unit (ReLU) activation function is used to identify features; the pooling layer reduces the number of features through the maximum pooling operation, retaining only the most important features; the flattening layer converts the multi-dimensional feature map into a one-dimensional array; and the output layer generates the final result. After the model is built, hyperopt is used for hyperparameter optimization to select the best parameter combination.

2.3. DC-CNN

The Dual-channel Convolutional Neural Networks with Attention-pooling (DC-CNN) model effectively improves the accuracy of fake news detection by combining multi-channel convolutional neural networks and attention pool optimization mechanisms [12]. The embedding layer of this model uses the DWtext method to generate word embeddings and generates context-related word vectors through word segmentation and data cleaning. In the process of data clarification, DWtext can filter out irrelevant information and retain useful classification features. DWtext ensures that words with the same context have similar semantics by predicting context word vectors. This method can also handle new words and derivative words at the same time; the convolution layer uses a variety of convolution kernels to extract text features and capture local dependencies between adjacent words; the dual-channel pooling layer extracts local and global features through Max-pooling and Attention-pooling respectively, among which the Max-pooling layer: is mainly used to extract local features, reduce redundant features, and improve the robustness of the model; the Attention-pooling layer uses a multi-head attention mechanism to capture long-distance dependencies and enhance the learning of global semantics; the classifier inputs the extracted features into the fully connected layer for classification and outputs the fake news detection results.

2.4. CNN-LSTM

This model is a hybrid neural network that combines the two basic models of CNN and RNN [13]. The model is mainly divided into five parts. The embedding layer converts the input news title and topic into word vectors and converts each word into a 100-dimensional vector. The number of input features is 5, 000, and the output is a 5, 000×100 matrix; the convolution layer extracts local features of the input text and uses 64 filters of different sizes for convolution operations; the maximum pooling layer reduces the feature dimension and retains important features; the LSTM layer processes sequence data and captures long-distance dependencies; the fully connected layer uses the softmax activation function for multi-classification output. In order to improve the efficiency and performance of the model, this model uses two feature selection and dimensionality reduction methods, Principal Component Analysis (PCA) and chi-square test (Chi-Square).

3. Discussion

Based on the above and similar studies, the identification of fake news has made great progress due to the introduction of deep learning. However, there are still some shortcomings and challenges in the current research.

The important thing to consider is the scalability of the model dataset. Currently, many fake news identification models have good performance on the given small data. However, whether the accuracy can reach the same level as that of the small dataset when migrating to a larger dataset remains to be

DOI: 10.54254/2755-2721/87/20241567

verified. Model developers often consider migrating the model to a larger dataset as a future work direction [11].

Similar, the models generally lack cross-language and cross-cultural generalization capabilities. When training the model, it is often only trained on texts in a single language, mainly English. Due to differences in culture and writing habits, a model trained in one language is difficult to effectively recognize articles in another language. In fact, due to the characteristics of some languages, such as Chinese and Japanese, there are no white spaces between words [14], so some models trained in non-language languages cannot be applied at all.

At the same time, model training speed is also a very thorny issue. Taking the CNN-LSTM model mentioned as an example, it takes 3 hours to train a training set of 50,000 samples based on 2 Intel Xeon 8-core 2.4GHz processors and 32GB DDR4 memory [11]. 3 hours of training time are a reasonable time for this task and model, but considering the training configuration, the training time of a machine with ordinary configuration will be longer. For actual application scenarios, this time may still need to be optimized.

Based on the above problems, a possible solution is to use transfer training to train new models based on existing related tasks, reduce the need for large-scale labeled data in the target domain, and improve the performance of the model in the target domain by adjusting the differences in features and data distribution between the original domain and the target domain [15, 16]. In the work of fake news prediction, for models of language style and sentiment analysis, transfer learning methods can be used to resolve differences between different data by reweighting instances in the source domain or discovering potential common feature spaces, so as to achieve the goal of not having to completely retrain the model.

Another possible method is to conduct model training based on distributed computing platforms, such as spark. Spark has the characteristics of memory computing, distributed computing, integration, etc. It can rely on memory to process data in parallel on the cluster, which can significantly improve the computing speed [17]. Spark also provides a wealth of high-level libraries (such as SparkSQL, GraphX), which simplify data preprocessing, feature extraction and other processes.

4. Conclusion

This work summarizes the recent innovations and applications of deep learning in fake news identification. This article focuses on three deep learning models used in fake news identification: OPCNN-FAKE, DC-CNN, and CNN-LSTM. Compared with general machine learning, they have better prediction accuracy. At the same time, in the field of fake news prediction, the scalability and training speed of the model are issues worthy of attention. To solve these two problems, this paper proposes two possible solutions: transfer learning and distributed computing, which can be considered the effective solutions for these issues.

References

[1] Watson C A 2018 Information literacy in a fake/false news world: An overview of the characteristics of fake news and its historical development International Journal of Legal Information vol 46 (Cambridge: Cambridge University Press) p 93-96

[2] Posetti J & Matthews A 2018 A short guide to the history of ‘fake news’ and disinformation International Center for Journalists vol 7 2018-07

[3] Moscadelli A, Albora G, Biamonte M A et al. 2020 Fake news and covid-19 in Italy: Results of a quantitative observational study International journal of environmental research and public health vol 17 (Basel: MDPI) p 5850

[4] Shu K, Sliva A, Wang S et al. 2017 Fake news detection on social media: A data mining perspective ACM SIGKDD explorations newsletter vol 19 (New York: ACM) p 22-36

[5] Mridha M F, Keya A J, Hamid M A et al. 2021 A comprehensive review on fake news detection with deep learning IEEE access vol 9 (Piscataway: IEEE) p 156151-156170

DOI: 10.54254/2755-2721/87/20241567

[6] Xu W, Wu J, Liu Q et al. 2022 Evidence-aware fake news detection with graph neural networks Proceedings of the ACM Web Conference 2022 (New York: ACM) pp 2501-10
[7] Altheneyan A & Alhadlaq A 2023 Big data ML-based fake news detection using distributed learning IEEE Access vol 11 (Piscataway: IEEE) p 29447-29463
[8] Sheng Q, Zhang X, Cao J et al. 2021 Integrating pattern-and fact-based fake news detection via model preference learning Proceedings of the 30th ACM international conference on information & knowledge management (New York: ACM) pp 1640-50
[9] LeCun Y, Bengio Y & Hinton G 2015 Deep learning nature vol 521 (London: Nature Publishing Group) p 436-444
[10] Agatonovic-Kustrin S & Beresford R 2000 Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research Journal of pharmaceutical and biomedical analysis vol 22 (Amsterdam: Elsevier) p 717-727
[11] Saleh H, Alharbi A & Alsamhi S H 2021 OPCNN-FAKE: Optimized convolutional neural network for fake news detection IEEE Access vol 9 (Piscataway: IEEE) p 129471-129489
[12] Ma K, Tang C, Zhang W et al. 2023 DC-CNN: Dual-channel Convolutional Neural Networks with attention-pooling for fake news detection Applied Intelligence vol 53 (Dordrecht: Springer) p 8354-8369
[13] Umer M, Imtiaz Z, Ullah S et al. 2020 Fake news stance detection using deep learning architecture (CNN-LSTM) IEEE Access vol 8 (Piscataway: IEEE) p 156695-156706
[14] Névéol A, Dalianis H, Velupillai S et al. 2018 Clinical natural language processing in languages other than English: opportunities and challenges Journal of biomedical semantics vol 9 (London: BioMed Central) p 1-13
[15] Weiss K, Khoshgoftaar T M & Wang D D 2016 A survey of transfer learning Journal of Big data vol 3 (New York: SpringerOpen) p 1-40
[16] Qiu Y, Hui Y, Zhao P, Wang M, Guo S, Dai B, Dou J, Bhattacharya S, Yu J 2024 The employment of domain adaptation strategy for improving the applicability of neural network-based coke quality prediction for smart cokemaking process. Fuel. 372:132162.
[17] Meng X, Bradley J, Yavuz B et al. 2016 Mllib: Machine learning in apache spark Journal of Machine Learning Research vol 17 (Cambridge: MIT Press) p 1-7

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241539

An analysis of risk control in the financial sector using big data technology

Dalong Lin

School of Computer Science, Guangdong University of Petrochemical Technology, Maoming, Guangdong, 525000, China

893245507@qq.com

Abstract. The rapid development of information technology has ushered in a new era of big data in the financial industry, which has provided financial institutions with a plethora of novel tools for risk management, while also playing a pivotal role in the industry's risk control landscape. This paper analyzes the public data of existing financial institutions, combines relevant literature, and employs big data technology to assess and identify risks, which aims to explore how big data technology can improve the risk management ability of the financial industry. In addition, the paper examines the challenges and problems faced by big data technology in the application of financial risk control, including data privacy and security, technology costs, and the demand for specialized talents. The research demonstrates that big data technology improves the speed and accuracy of risk identification, but also points out the challenges of data privacy and security management. Future risk control models need to better integrate big data analytics with traditional risk management methods, thus promoting the further development and application of big data technology in the field of risk control.

Keywords: Big Data, Internet Finance, Risk Challenges, Financial Industry.

1. Introduction

In the 21st century, driven by the trend of informationization, big data technology is reshaping the ecological landscape of the modern financial management industry. Particularly in risk management, the application of big data technology has significantly enhanced risk identification precision and evaluation effectiveness, while also fostering profound innovation and personalized development within financial services. However, alongside the widespread adoption of big data technology, a range of complex and severe challenges have emerged. For instance, issues surrounding data security and privacy protection have become prominent concerns, while talent shortages and skills gaps pose potential threats to the healthy and stable growth of the financial industry. Therefore, it holds immense theoretical and practical significance to thoroughly discuss both the impact and challenges posed by big data technology on risk management within the financial industry. This paper employs a literature review and data analysis to assess and identify risks, with the objective of exploring how big data technology can enhance the risk management capability of the financial industry. Furthermore, this paper examines the challenges and problems encountered by big data technology in the application of financial risk control. The results indicate that big data technology enhances the speed and accuracy of risk identification while also identifying challenges in data privacy and security management. Future

DOI: 10.54254/2755-2721/87/20241539

risk control models must integrate big data analytics with traditional risk management methods to further develop and apply big data technologies in the field of risk control.

2. Overview of Big Data Development and Applications

Big data, known for its massive data volume, diverse data types, fast processing speed and profound value potential, is also referred to as 4V. In the financial world, big data has become ubiquitous, giving financial companies unprecedented risk insight and assessment power. These data are not only numerous, but also cover a variety of types, such as transaction records, user behavior, market dynamics, etc., which together constitute a comprehensive risk profile of financial institutions

2.1. Main Application of Big Data

Nowadays, big data technology has permeated various industries and sectors, emerging as a pivotal force driving social progress and development. In the financial domain, the application of big data technology is extensive and profound, offering robust support for risk management, business innovation, and customer service within financial institutions. The manufacturing industry leverages big data to optimize production processes and achieve intelligent manufacturing. Transportation systems employ big data analytics to alleviate congestion and enhance efficiency. The telecom sector optimizes services through customer data analysis [1]. Educational institutions also utilize learning data analysis to enhance teaching quality. As technology advances further, big data will be more deeply integrated into these fields, underscoring the increasing significance of ensuring data security and privacy protection. These applications not only boost industry efficiency but also provide substantial impetus for societal progress [1].

2.2. Impact of Big Data on the Financial Sector

The profound impact of Big Data on the financial industry cannot be overstated. The utilization of big data technology has revolutionized the operations of financial institutions, significantly enhancing their capacity and efficiency in risk management. Furthermore, it empowers these institutions to conduct more precise risk assessments and develop more effective management policies. Additionally, the application of big data technology has fostered innovation in financial services by promoting personalized offerings and products, thereby enabling financial institutions to cater to diverse customer needs and risk appetites effectively. These advancements not only benefit financial institutions but also provide customers with tailored financial solutions.

3. Risk Management in Internet Finance and its Associated Risks

3.1. Definition and Classification of Internet Finance

Internet finance, as a result of the deep integration between finance and technology, is progressively emerging as a pivotal component within the financial industry due to its inherent attributes of convenience, efficiency, and inclusivity. It encompasses various sub-domains such as online lending, Internet payment systems, Internet-based insurance services, and Internet-driven fund sales; thereby presenting novel avenues for the advancement of traditional financial services.

3.2. Role of Internet Finance in Society

Internet finance plays a pivotal role in contemporary society, exerting profound influence on economic development, social progress, and individuals' lives with its distinctive advantages and innovative approach. Firstly, Internet finance offers more convenient and cost-effective financing channels for small and medium-sized enterprises (SMEs) as well as individuals. Traditional financial institutions often exhibit caution towards the financing needs of SMEs and individuals due to factors such as cost and risk [2]. Leveraging big data, cloud computing, and other technological means, internet financial platforms can swiftly and accurately assess borrowers' credit status and repayment capacity, thereby providing more flexible and personalized financing services [2]. Secondly, the development of inclusive

DOI: 10.54254/2755-2721/87/20241539

finance has been aided by online financing. Internet finance is a vital tool in accomplishing the goal of inclusive finance, which is to make financial services more accessible to a larger segment of society. The accessibility gap to financial services can be closed by providing low-income and remote communities with the same financial services options as metropolitan populations using online financial platforms. Additionally, internet finance has stimulated innovation within the financial market's development. By introducing novel business models, technical approaches, and service concepts; internet-based financial platforms have disrupted traditional financial institutions' monopoly position while promoting competition within the market along with reform initiatives. This transformation not only enhances efficiency levels alongside service quality but also infuses new vigor into the overall landscape of the financial market.

3.3. Risk Control Strategies for Internet Finance

However, the rapid development of Internet finance has brought about increasingly prominent risks and challenges. In order to effectively address these risks, Internet financial enterprises have implemented a series of risk control strategies. To cope with diversified risks, comprehensive risk control strategies have been adopted by these enterprises. These encompass various dimensions such as legal compliance, technical security, user education, and the utilization of big data technology for user behavior analysis and transaction monitoring. Through the implementation of real-name authentication, credit scoring systems, black-gray list management, and other measures, multi-level risk management systems have been established [3]. Additionally, regular strategy monitoring and review ensure that risk control measures are updated in response to market changes to safeguard business operations' robustness and client asset safety. This comprehensive risk control framework not only ensures the healthy development of Internet finance but also enhances trustworthiness and reliability within the entire industry.

4. Emerging Trends in Risk Control

In today's digital age, the rapid development of Internet finance has made risk control one of the important challenges faced by financial institutions. In order to meet the increasingly complex and changeable risk environment, the integration of deep algorithms, big data and machine learning technology has become a new trend in the field of risk control. The combined application of these techniques in risk control and the changes they bring are discussed in detail below.

4.1. Advanced Applications of Depth Algorithms in Risk Control

The advent of powerful learning and processing capabilities, coupled with the advent of deep algorithms, has brought a novel perspective to the field of risk control. The construction of deep neural network models enables financial institutions to conduct in-depth analysis of complex features and patterns in user data, thereby facilitating more accurate risk identification and prediction. In the field of credit assessment, traditional scoring models frequently rely on credit data and simple statistical methods. However, with the advent of big data technology, financial institutions have gained access to a greater variety of user data, including social network data and consumer behavior data. Credit scoring models are constructed using deep neural networks (e.g., CNNs or RNNs), which are trained with voluminous data to discern intricate interrelationships between user information profiles.

4.2. Deep Integration of Big Data and Machine Learning

Big data provides rich data resources for machine learning, and machine learning algorithms can dig out potential risk rules from massive data. In the field of risk control, the deep integration of big data and machine learning has realized real-time monitoring, automatic identification and early warning of risks. Fraud is a common risk in payment platforms. In order to effectively identify fraud, a payment platform has adopted a fraud detection system based on machine learning. By monitoring users' transaction data and behavior patterns in real time, the system uses support vector machine (SVM), Random Forest and other algorithms to build fraud prediction models. Once the system identifies an abnormal transaction or

DOI: 10.54254/2755-2721/87/20241539

behavior pattern, it will immediately trigger an early warning mechanism and take appropriate risk control measures. In addition, the system can dynamically adjust risk management strategies according to market changes and risk challenges to ensure the safety of funds.

4.3. Integrated Applications of Deep Algorithms, Big Data and Machine Learning

In the field of risk control, the integrated application of deep algorithms, big data and machine learning can achieve more accurate and efficient risk control. By building deep learning models, financial institutions can deeply mine complex features and patterns in user data; Big data provides rich data resources for machine learning, enabling machine learning algorithms to identify and predict risks more accurately. At the same time, big data and machine learning technology can also realize real-time monitoring and dynamic adjustment of risks, ensuring that financial institutions can timely respond to market changes and risk challenges [4].

4.3.1. Comprehensive Risk Control Platform. To improve the risk control ability, a comprehensive risk control platform can be built, which integrates a variety of technologies such as deep learning, big data, and machine learning, and realizes the comprehensive monitoring and management of various types of risks. Through deep learning models, the platform can automatically learn and extract complex features from user data; At the same time, big data provides platforms with rich data sources that allow machine learning algorithms to identify and predict risks more accurately. In addition, the platform also has real-time monitoring and dynamic adjustment capabilities, which can adjust risk management strategies in a timely manner according to market changes and risk challenges. This comprehensive application not only improves the accuracy and efficiency of risk control, but also realizes the intelligence and individuation of risk management. In summary, the integration of deep algorithms, big data and machine learning in the field of risk control provides a new solution for financial institutions.

4.3.2. Innovative Practice of Data mining and Transmission in Risk Control. Data mining and transmission play a crucial role in the risk control of Internet finance. Through in-depth mining and analysis of user data, financial institutions can find valuable information hidden behind the data, such as the user's credit status, consumption habits, investment preferences, etc. This information not only helps financial institutions to develop more accurate risk management strategies, but also provides decision-making support for financial institutions. At the same time, efficient data transmission mechanisms ensure that risk control measures can be implemented in a timely and effective manner. By establishing a sound data sharing and exchange platform, financial institutions can realize real-time data sharing and collaborative work, and improve the efficiency and accuracy of risk control work.

4.3.3. Core Role of data Processing and Analysis in Risk Control. Data processing and analysis is one of the core applications of big data technology in Internet financial risk control. In the face of massive user data, financial institutions need to extract valuable information through efficient data processing and analysis technology, which includes data cleaning, integration, analysis and visualization. Through data processing and analysis, financial institutions can more comprehensively understand the user's behavior characteristics, credit status and risk level, and provide strong support for risk decision-making. At the same time, data processing and analysis technology can also help financial institutions find potential risk points and trends, formulate risk response strategies in advance, and reduce risk losses.

5. Challenges of Big Data Technologies for Risk Control and Management

5.1. Data Security and Privacy Protection

With the wide application of big data technology in risk control and management, data security and privacy protection issues have become increasingly prominent. Given the substantial volume of sensitive data involved in risk control management, including user identity information and transaction records, ensuring the security and privacy of these data sets has become a significant challenge for

DOI: 10.54254/2755-2721/87/20241539

financial institutions. Once the data is leaked or abused, it may lead to serious consequences, such as damage to the rights and interests of users and damage to the reputation of financial institutions. Therefore, financial institutions need to strengthen the research and development and application of data security management and privacy protection technology to ensure data security in the process of risk control management.

5.2. Technology Update and Talent Shortage

The rapid development of big data technology has placed higher requirements for technical update and talent training of financial institutions. However, there is a relative shortage of professionals in the field of big data in the market at present, which makes financial institutions face certain difficulties in introducing and applying big data technology. In order to meet this challenge, financial institutions need to increase personnel training and introduction efforts to improve the level of big data technology and application capabilities of employees. Meanwhile, financial institutions should also strengthen cooperation with universities and research institutions to jointly promote the development and application of big data technology. [5].

5.3. Regulatory and Compliance Challenges

The application of big data technology in risk control and management is becoming more and more extensive, and relevant regulatory and compliance requirements are becoming increasingly stringent. Financial institutions need to carry out risk control management under the premise of complying with laws and regulations to ensure data compliance and legitimacy. At the same time, financial institutions also need to strengthen communication and cooperation with regulators to jointly promote the healthy development of big data technology in the field of risk control and management. In addition, financial institutions need to be aware of international data protection regulations and standards to ensure compliance when doing business globally [6].

6. Conclusion

To sum up, big data technology has brought far-reaching impacts and challenges to the risk control and management of the financial industry. Through the deep integration and innovative practice of deep algorithms, big data, machine learning and other technologies, financial institutions can better cope with risk challenges and improve risk management. However, issues such as data security and privacy protection, technological updates and talent shortages, and regulatory and compliance challenges cannot be ignored.

Looking ahead, with the continuous progress of technology and the improvement of regulations, the application of big data technology in the risk control management of the financial industry will be more extensive and in-depth. Financial institutions will be able to use more advanced data mining and analysis techniques to achieve more accurate identification and assessment of risks. At the same time, with the continuous development of emerging technologies such as artificial intelligence and blockchain, big data technology will integrate with these technologies to jointly promote the innovation and upgrade of risk control management in the financial industry. with the acceleration of globalization and digital transformation, financial institutions also need to strengthen international cooperation to jointly address cross-border risk challenges and achieve globalization and synergy in risk management and control. And Internet financial enterprises should make full use of big data software systems to conduct comprehensive and accurate analysis of various investment portfolios, so as to improve the quality of Internet financial investment. Furthermore, Internet financial enterprises should employ a comprehensive range of view-based and automated models to effectively identify and predict Internet financial risks, and implement targeted risk prevention measures on this basis.

DOI: 10.54254/2755-2721/87/20241539

References

[1] Chen, Y.Y. (2022) Research on Recommendation Method andPrivacy Protection for New N EMT Subject Selection. Hebei Normal University. https://www.doc88.com/p-1176155430 3819.html

[2] DOC88.COM. (2016) Research on the Innovation of Financing Mode of Small and Mediu m-sized Coal Mining Enterprises. https://www.doc88.com/p-9912711950757.html?s=rel&i d=1

[3] Yuan, S.H. (2023) A Study of Artificial Intelligence Elements and Implications for China. Science & Technology Industry of China. 2023(07): 64-66.

[4] SXWORKER.COM. (2023) Artificial Intelligence for Industrial Manufacturing: It's Just Beginning. http://paper.sxworker.com/xpaper/news/1461/6705/54688-1.shtml

[5] Zhang, J.L. (2019) Cost and Management Optimization of Management Trainee Programs in Foreign Financial Companies. Shanghai University of Finance and Economics (SUFE)

[6] Luo, G.H. (2022) Research on Internet Financial Risk Prevention Strategies in the Era of Big Data. Trade Fair Economy. 2022(21): 059-061.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241573

Detection of network false information based on artificial intelligence models

Haoxi Mao

Computer Science, Shanxi Agricultural University, Jinzhong, Shanxi, 030801, China

quxin@ldy.edu.rs

Abstract. There is often some false information in social platforms to mislead public opinion. Due to the rapid development of the Internet, the spread of false information on the Internet has become easier, which has brought many losses to people's economy and life. In this paper, the relevant research on false information based on bidirectional convolutional networks is analyzed, and the method is divided into four stages: data preprocessing, model architecture, training process and prediction process. Then, the relevant research on rumor detection by propagation tree kernel model are analyzed. Finally, this paper delineates a comprehensive framework that amalgamates enhanced Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and a hybridized Black Widow Optimization (BWO) with Moth Optimization Algorithm (MOA) (referred to as HM-BWO) for the accurate detection of network-based false information. This paper analyzes and discusses the challenges of poor universality and insufficient detection speed encountered by the current rumor detection research and puts forward the idea of introducing migration model to solve the problem of poor universality and Spark to solve the problem of insufficient detection speed. This article provides a good overview of the field of online disinformation.

Keywords: machine learning, natural language process, rumor detection.

1. Introduction

Network false information refers to information presented in the form of words, images, audio and video or symbols in the carrier of modern information networks, which is inconsistent with the facts or fabricated, disrupts the political, social, economic and other order, or infringes on the legitimate rights and interests of others. In social media, there is usually some false information to mislead the public opinion, so as to achieve the purpose of obtaining economic benefits or achieving certain political goals. With the progression of the internet era, the substantial rise in the quantity of internet users, and the dissemination of online social networking practices, the propagation of misinformation on the web is more facile than that via conventional news mediums such as newspapers and radio. The nominal expense associated with sustaining social media platforms, coupled with their user-friendly interfaces, exacerbates this developing trend. By comparing the depth, size, maximum breadth and structural virality of the cascade of false information and true information forwarding, it can be observed that false information disseminates significantly more extensively, rapidly, deeply, broadly, and structurally perniciously than true information [1]. The losses caused by false information are huge. For example, BBC News reported on August 12, 2020, that a new study showed that in the first three months of this

DOI: 10.54254/2755-2721/87/20241573

year alone, false information about the novel coronavirus on social media had led to at least 800 deaths and about 5,800 hospitalizations worldwide [2]. Therefore, research on the detection of false information on the Internet is very necessary. In the face of the massive network information generated every day, traditional manual detection methods cannot accurately distinguish false information, while artificial intelligence has a strong ability to extract text image information feature values and prediction, and can quickly identify false information on the network and reduce the harm brought by false information on the network.

Artificial intelligence has made great progress in recent years. It has a variety of algorithms, such as logistic regression model, decision tree, naive Bayes function, convolutional neural network, etc., which have been applied in finance, medical treatment, games, risk assessment, data analysis and other fields. In journalism, there have been many people using artificial intelligence technology to do related research, and the detection of false information on the Internet is an important direction of journalism. At present, there are many researchers using the relevant model of artificial intelligence to study the detection of false information on the Internet. For instance, the Binary Graph Convolutional Network (Bi-GCN) model delves into these two aspects by employing a concurrent top-down and bottom-up approach to rumors' dissemination. The system employs the Graph Convolutional Network (GCN) in conjunction with a top-down directed graph to elucidate the patterns of rumor propagation, and conversely, integrates the GCN with the inverse graph of rumor propagation to encapsulate the architecture of its dissemination [3]. Furthermore, certain studies have embraced the non-linear structural characteristics of the propagation tree, integrating them with linear features for the purpose of rumor classification [4]. In addition to applications in machine learning, deep learning models are also applicable in pertinent research concerning the detection of false information on the internet. For instance, graph neural networks are employed to acquire the representation of user relevance from the binary graph encapsulating the correlation between users and source tweets [5], and the representation of information propagation using tree structure. This paper then combines the representations learned from these two modules to classify the rumors.

The present study endeavors to furnish a thorough summary of the artificial intelligence mechanisms employed in the detection of misinformation across the Internet. It is mainly composed of four parts and the rest is organized as follows. First, in the second part, this review will elaborate on the methods used to detect false information on the Internet in detail. In the third part, the current status and development of the current network false information detection and the challenges it faces. In the last part, I will make a summary of the whole article.

2. Method

2.1. Introduction of machine learning

Network false information detection based on machine learning mainly includes data collection, data preprocessing, feature extraction, selection of learning model, training model, model evaluation, model optimization and practical application shown in Figure 1.

Data collection: Collect large amounts of textual data containing true information and rumors.

Data preprocessing: the text is cleaned, the word is divided, the word is stopped and so on, and the text is transformed into a form suitable for model processing.

Feature extraction: Extract meaningful features from preprocessed text, such as word frequency, part of speech, semantic features, etc.

Select learning model: It is imperative to choose an appropriate machine learning model that aligns with the distinct attributes of the problem at hand.

Training model: Utilize the annotated training data to refine the model.

Model evaluation:Utilize an exclusive test dataset to assess the efficacy of the model.

Model optimization: Adjust and optimize the model based on the evaluation results, such as adjusting parameters, trying different models, or combining multiple models.

Practical application: The optimized model is applied to new text data for rumor detection.

DOI: 10.54254/2755-2721/87/20241573

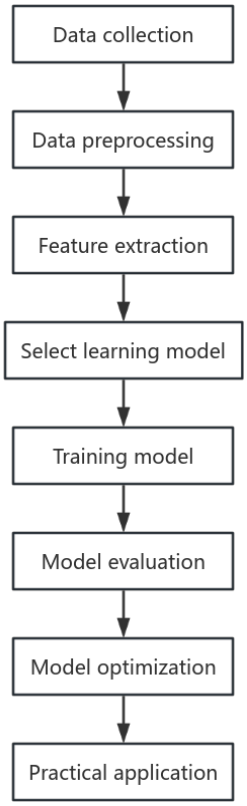

Figure 1. Flow chart of network false information detection based on machine learning (Photo/Picture credit: Original).

2.2. Machine learning models

2.2.1. Bi - GCN

The authors introduce a novel rumor detection approach utilizing a Bidirectional Graph Convolutional Network (GCN), designed to identify rumors on social media platforms. This method integrates the attributes of both rumor propagation and diffusion, conducting an analytical investigation of rumor characteristics through traversal in both upward and downward propagation directions. The methodology is primarily composed of four sequential phases: data preprocessing, architectural design of the model, the training regimen, and the subsequent prediction phase. During the data preprocessing stage, the researcher employed a trio of datasets: Weibo, Twitter15, and Twitter16. These data sets contain information such as users, posts, retweets and reply relationships. The authors extracted the TF-IDF values of the posts as features, and built a propagation structure based on the forward and reply relationships. The authors then divided the data into training sets, verification sets and test sets based on the rumor tags. Using DropEdge during the training phase to avoid overfitting problems. The researchers employed a bidirectional Bi-GCN architecture for the identification of rumors within social media platforms.This architectural design integrates a Top-Down graph Convolutional network (TD-GCN) with a Bottom-Up graph Convolutional network (BU-GCN). The TD-GCN facilitates the dissemination of rumors by enhancing the transmission of information from parent nodes to their descendant nodes. Conversely, the BU-GCN encapsulates rumor propagation by aggregating information from descendant nodes to their parent node. The model splices the output of TD-GCN and BU-GCN to obtain the final rumor detection result [3].

DOI: 10.54254/2755-2721/87/20241573

2.2.2. Propagation tree kernel model

The author uses the Propagation Tree Kernel (PTK) and the Context-Sensitive Propagation Tree Kernel (cPTK) to detect rumors.

Two datasets, Twitter15 and Twitter16, were used in the experiment. These two datasets contain a large number of source tweets and their spread structure, and have been labeled as either rumors or non-rumors.The efficacy and preeminence of the proposed propagation tree kernel model, as well as the context-sensitive propagation tree kernel model, have been empirically substantiated through experimental validation on the respective datasets.

In the course of the experiment, the author delineates the dissemination of every individual tweet as a hierarchical tree framework. Herein, the root node corresponds to the originating tweet, while the leaf nodes signify the audience's responses to the post. Furthermore, the directed edges encapsulate the inter-node responsive relationships.Subsequently, the higher-order structure of diverse rumor types is delineated through the computation of the likeness between their propagation trees.Certainly, the Pattern Tree Kernel (PTK) or its compressed variant (cPTK) is utilized to measure the similarity within these propagation trees. Subsequently, these trees are incorporated as features into a kernel-based Support Vector Machine (SVM) classifier for the intent of classification. Within the confines of a multi-classification endeavor, the one-to-many classification strategy is adopted, wherein the category garnering the highest likelihood is nominated as the predictive outcome [4].

2.2.3. ICNN

In the current research, they present a novel hybrid deep learning architecture for the detection of fake news. This framework integrates Enhanced Convolutional Neural Networks (ECNN), Long Short-term Memory Networks (LSTM), and a hybridized approach incorporating the Black Widow Optimization (BWO) and Moth Optimization (MO) algorithms to facilitate the automated detection and categorization of fraudulent news content prevalent on social media platforms. The data utilized in the analysis is sourced from the Fake News Challenges (FNC) repository, as well as the KDnuggets and ISOT datasets.

In this experiment, the authors used an improved convolutional neural network structure called ICNN. The Integrated Convolutional and Recurrent Neural Network (ICNN) harnesses the merits of both convolutional and recurrent neural networks, facilitating the concurrent processing of textual spatial and temporal information. The architecture of the Integrated Convolutional Neural Network (ICNN) comprises an input layer, a convolutional layer, a pooling layer, a recurrent layer, a fully connected layer, and an output layer. The input layer is tasked with processing textual data, whereas the convolutional and pooling layers are responsible for the extraction of textual features. The recurrent layer applies recurrent processing on the extracted features, thereby enhancing the network's ability to model temporal dependencies [5].

3. Discussion

3.1. Limitations and challenges

3.1.1. Generality

Generally, the generality of the model is strongly related to the training set used in the training model. For example, the data set of a social platform is used to train the false information detection model, and the trained model is used to detect false information on another platform, which will fail to accurately detect false information. Because their data distribution is different, such as video data, text data, user comment data etc, there are great differences, resulting in poor universality of detection models. However, due to the poor universality of the detection model, it is required to train the model separately for different models and different language countries, resulting in a great increase in cost. Therefore, how to train the detection model with certain universality or easy migration through the field of easy data collection for cross-domain, cross-platform and cross-source information detection is a huge challenge that cannot be avoided in the application of false detection technology.

DOI: 10.54254/2755-2721/87/20241573

3.1.2. Speed

In the real-time application environment, information detection is faced with massive data flow, and false information spreads viral far faster than true information. For the release of some information, time sensitivity is very important. The core of human curiosity and the pursuit of novelty constitutes the foundational factor that ascertains the timeliness value of news content. Daily, a vast quantity of information traverses social media platforms, and there is a prevalent preference among internet users for the most recent updates, which in turn intensifies the necessity for the immediate identification and verification of misinformation. At the same time, there are great requirements for the speed of detecting false information. Training models to detect faster is a big challenge.

3.2. Future prospects

3.2.1. Transfer learning

Transfer learning is a kind of learning method in machine learning. Transfer learning facilitates the application of established learning models within diverse, yet interconnected, environments. In conventional machine learning approaches [6-8], the models lack sufficient flexibility, resulting in suboptimal performance when addressing variations in data distribution, dimensionality, and model output alterations. Transfer learning mitigates these constraints by incorporating knowledge from the source domain, thereby enhancing the modeling process under varying conditions of data distribution, feature dimensions, and model output dynamics [6]. The combination of transfer learning and false information detection model can effectively solve the problem of poor generalization of the model.

3.2.2. Spark

Apache Spark is an expedient big data processing engine that excels in distributed computing environments, enabling the efficient manipulation of massive datasets across clusters. This technology has garnered significant popularity in recent times. This framework is poised to supersede Hadoop, as depicted in Figure 2. Its primary benefits include rapid processing, user-friendliness, and exceptional adaptability. Its paramount feature is its expediency, boasting speeds that are 100 times faster in memory and 10 times faster on disk when compared to HadoopMapReduce [9, 10].

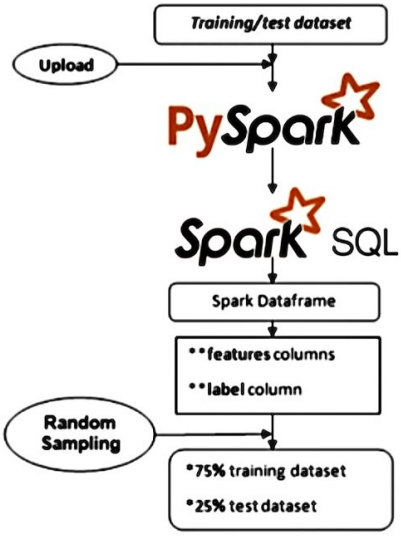

Figure 2. Structure based on spark SQL training and testing data sets [7].

In recent years, many studies have used Apache Spark to conduct data analysis and processing in the data preprocessing stage of network false information detection, but few have combined it with model

DOI: 10.54254/2755-2721/87/20241573

training. In the future, it is very promising to combine the model training of rumor Apache Spark and network false information detection to improve the speed of training model and model optimization.

4. Conclusion

This manuscript has provided an exhaustive summary of the methodologies employed in the detection of misinformation on the internet. In this paper, researches on the detection of network false information by Bi-GCN, Propagation tree kernel model and ICNN are investigated respectively, and their specific processes and algorithms are explained in detail. After discussion and analysis, it could be found that in the field of network false information detection, there are mainly two problems: poor universality of detection model and insufficient detection speed. To solve these two problems, this paper proposed a solution that uses Spark platform to solve the insufficient speed of model detection and uses transfer learning method to solve the poor universality of detection model. In the future, the further study plans to apply the proposed hypothesis to serve the research of network false information detection.

References

[1] Vosoughi S, Roy D & Aral S 2018 The spread of true and false news online Science vol 359 (6380) pp 1146-1151
[2] Sina Science and Technology 2020 Eating garlic to protect against COVID-19? Study: Nearly 5,800 people were hospitalized worldwide because of misinformation related to COVID-19, at least 800 deaths [EB/OL]
[3] Bian T, Xiao X, Xu T, Zhao P, Huang W, Rong Y & Huang J 2020 Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)
[4] Hamidian S, Diab MT 2019 Rumor detection and classification for twitter data. arXiv preprint arXiv:1912.08926
[5] Narang P, Singh AV & Monga H 2022 Hybrid Metaheuristic Approach for Detection of Fake News on Social Media International Journal of Performability Engineering vol 18 no 6 June pp 434-443
[6] Pan SJ & Yang Q 2010 A Survey on Transfer Learning IEEE Transactions on Knowledge and Data Engineering vol 22 no 10 pp 1345-1359
[7] Qiu Y, Hui Y, Zhao P, Wang M, Guo S, Dai B, Dou J, Bhattacharya S & Yu J 2024 The employment of domain adaptation strategy for improving the applicability of neural network-based coke quality prediction for smart cokemaking process Fuel Sep 15 vol 372 p 132162
[8] Ma Y, Chen S, Ermon S & Lobell DB 2024 Transfer learning in environmental remote sensing Remote Sensing of Environment Feb 1 vol 301 p 113924
[9] Madani Y, Erritali M & Bouikhalene B 2021 Fake News Detection Approach Using Parallel Predictive Models and Spark to Avoid Misinformation Related to Covid-19 Epidemic In: Gherabi N & Kacprzyk J (eds) Intelligent Systems in Big Data, Semantic Web and Machine Learning Advances in Intelligent Systems and Computing vol 1344 Springer, Cham
[10] Öztürk MM 2024 Tuning parameters of Apache Spark with Gauss–Pareto-based multi-objective optimization Knowledge and Information Systems Feb vol 66 (2) pp 1065-1090

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241611

Research on macroeconomic indicators and stock market correlation analysis based on machine learning

Haocheng Tian

Financial Engineering, University of Southern California, Los Angeles, 90007, China

haochengtian123@gmail.com

Abstract. The stock market is known as the barometer of the national economy, and macroeconomic factors have an important impact on the volatility of the stock market. Therefore, considering the impact of macroeconomic factors on the sustainability of stock market volatility will help to capture the time-varying characteristics of volatility persistence, so as to significantly improve the estimation and forecasting effect of volatility. In this work, the generalized autoregressive conditional heteroskedasticity-mixing data sampling (GARCH-MIDAS) model is adopted, which combines the advantages of the GARCH model in short-term volatility modeling and the advantages of MIDAS regression in integrating macroeconomic variables of different frequencies. The GARCH model provides an accurate depiction of intraday volatility in financial markets by capturing the dynamic nature of short-term volatility. MIDAS regression introduces long-term macroeconomic factors into volatility modeling by integrating macroeconomic data with different sampling frequencies, thus making up for the shortcomings of traditional measurement methods in frequency matching. By combining these two methods, the GARCH-MIDAS model can more comprehensively reflect the dynamic changes of stock market volatility, considering both the impact of short-term market volatility and the role of long-term macroeconomic factors, thus providing a more accurate and in-depth analytical tool for volatility prediction and risk management. The results show that the GARCH-MIDAS model can significantly improve the accuracy of volatility forecasting, and provide more reliable decision support for investors, policymakers and economists.

Keywords: Macroeconomic Indicators, Stock Market, Correlation Analysis, Long-short Term, Machine Learning.

1. Introduction

A healthy stock market is an important guarantee for the stable development of the national economy, and stock price fluctuations are one of the basic characteristics of the stock market. Investors can buy stocks when the stock price is low and sell the stock when the stock price is high, so as to achieve greater gains. The fluctuation of stock prices not only provides investors with profit opportunities, but also helps to realize the resource allocation function of the stock market. The huge and frequent volatility of the stock market not only affects the behavior of investors, but can also hinder the continued healthy development of the economy [1]. Therefore, studying the volatility of the stock market can help to better understand the operating laws and mechanisms of the stock market.

How to model and predict stock market volatility has become a hot issue for many scholars. For practitioners and regulators in the financial markets, it is essential to have accurate volatility forecasts.

DOI: 10.54254/2755-2721/87/20241611

A large number of scholars have conducted in-depth research on this purpose, building a series of models to predict volatility. Accurate volatility forecasting not only helps investors avoid market risks, but also helps regulators maintain market stability [2]. Considering the impact of macroeconomic factors on the sustainability of stock market volatility in the model will help to capture the time-varying characteristics of volatility more accurately, thereby significantly improving the effectiveness of the model in in-sample data fitting and out-of-sample forecasting [3].

The stock market is an important part of the development of the national economy and has three major functions: first, the financing function, commercial enterprises can issue stocks to raise funds in order to achieve rapid development; the second is the investment function, where investors can invest by buying and selling stocks; The third is the function of optimal allocation of resources, which transfers scarce resources from poor-performing enterprises to better-performing enterprises to promote the rational allocation of resources [4]. Therefore, the stable and healthy development of the stock market is crucial to the sustained and healthy development of the economy.

Changes in macroeconomic conditions and macroeconomic policies have an impact on the stock market. Investors can avoid investment risks in the stock market in a timely manner by collecting information on macroeconomic changes; Regulators are also able to adjust policies in response to changes in economic conditions [5]. Therefore, the study of macroeconomic conditions and stock market volatility prediction is of great significance for risk management, decision-making, economic development and the healthy operation of the stock market [6].

For market investors, they can adjust the scale and direction of investment according to the operation of the macroeconomy, so as to reasonably avoid risks and obtain maximum returns. For listed companies, macroeconomic conditions can affect stock price fluctuations, which in turn affect the size of the company's assets. Accurate prediction of stock market volatility can also reduce the company's financing cost and promote the sustainable and healthy development of the company [7]. As far as government supervision departments are concerned, they can guard against stock market risks, rationally carry out macro-prudential supervision, standardize the operation of the stock market, and make the stock market develop steadily and healthily according to the operation of the macroeconomy. Therefore, it is necessary to study the persistence of macroeconomic conditions and stock market volatility [8].

2. Related Work

A large number of studies on stock market volatility have shown that macroeconomic conditions are the main source of stock market volatility. Li et al. [9] examined the impact of monetary policy on U.S. and Canadian stock prices. However, there is a common drawback of existing research methods, that is, the data must have the same sampling frequency. For stock market data, we can get daily data or even higher frequency data, while for many macroeconomic variable data, we can usually only get monthly or quarterly data, and GDP data can only be obtained on a quarterly or annual basis. Traditional metrology methods require that the sampling frequency of variables in the model be consistent. To address this issue, many scholars have chosen to reduce the sampling frequency of stock market data to align it with that of macroeconomic variables.

Subsequently, Kim and Nelson [10] divided the volatility of the stock market into two parts (CR), one related to the economic cycle and the other independent of the economic cycle, and the results showed that the economic cycle affects the volatility of the stock market, exploring the relationship between macroeconomic conditions and stock market volatility. However, this approach leads to the loss of high-frequency effective information in the stock market, which in turn leads to errors in parameter estimation and volatility forecasting, making it impossible to fully assess the impact of macroeconomic information on stock market volatility. Therefore, studying how to build effective models between data with different sampling frequencies is the key to fully understanding the impact of macroeconomic conditions on stock market volatility.

In order to include variables with different sampling frequencies in the same model, Ghysels et al. [11] proposed a mixed data sampling (MIDAS) method. This method can make full use of existing information, so it has been widely used by many scholars. Engle et al. [12] utilized the generalized

DOI: 10.54254/2755-2721/87/20241611

autoregressive conditional heteroskedasticity (GARCH) model to divide volatility into long-term components and short-term components, and the long-term components are described by macroeconomic variables, so as to study the relationship between stock market volatility and macro fundamentals. The model not only solves the modeling problem of different frequency data, but also makes full use of the existing information to provide more accurate analysis results.

3. Methodologies

The relationship between macroeconomic indicators and stock market volatility is crucial for investors, policymakers, and economists. To analyze these complex relationships, we employ the Generalized Autoregressive Conditional Heteroskedasticity-Mixed Data Sampling (GARCH-MIDAS) model, which combines the advantages of the GARCH model in short-term volatility modeling and the advantages of mixed data sampling regression in integrating macroeconomic variables of different frequencies.

3.1. GARCH Model

The short-term volatility section uses the GARCH(1,1) model to capture the intraday volatility characteristics of the stock market. The GARCH (Generalized Autoregressive Conditional Heteroskedasticity) model is a model commonly used in financial time series analysis, which can effectively describe the phenomenon of volatility aggregation in financial markets. The calculation of the GARCH(1,1) model is shown in Equation 1.

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{1}$$

Where σ_t^2 is the conditional variance, which means the volatility of t at the current moment given past information. This is what we want to estimate through the model, reflecting the current level of uncertainty in the market. α_0 is a constant term. It indicates the fundamental level of volatility in the absence of other influencing factors. α_1 is the coefficient of the lag square residual, which represents the effect of the square of the previous period residuals on the current volatility. Specifically, ϵ_{t-1}^2 represents the square of the previous period residual, which captures the "shock effect" of volatility, which is how strongly the volatility reacts after a shock. β_1 is a coefficient of the lag variance, which represents the effect of the previous period's conditional variance on the current volatility. This is used to capture the "persistence effect" of volatility, which is the persistence characteristic of market volatility. Generally speaking, the greater the β_1, the more persistent the volatility is.

The advantage of the GARCH(1,1) model is that it is able to dynamically adjust the current conditional variance through past residuals and past variance, thus capturing the dynamic nature of volatility in financial markets. Specifically, when the market experiences large price movements, the square term of the residuals increases, resulting in an increase in the conditional variance; When the market is relatively stable, the conditional variance is mainly determined by the past conditional variance, reflecting the persistence of market fluctuations.

Through the maximum likelihood estimation method, we can estimate the parameters $\alpha_0, \alpha_1, \beta_1$ in the model to determine the dynamic change law of short-term volatility. Estimates of these parameters can help us understand the nature of market volatility and inform further volatility forecasting and risk management.

3.2. MIDAS Regression

The long-term volatility section is modeled using macroeconomic variables of varying frequencies. The mixed-frequency data sampling regression method allows us to introduce low-frequency macroeconomic variables into high-frequency financial time series models, so as to effectively capture the characteristics of long-term volatility changes. The calculation of the MIDAS regression model is shown in Equation 2:

$$\tau_t = \theta_0 + \theta_1 \sum_{k=0}^{K} w_k X_{t-k} \tag{2}$$

DOI: 10.54254/2755-2721/87/20241611

Where τ_t is the long-term volatility component, which reflects the level of volatility affected by macroeconomic variables. θ_0 and θ_1 are parameters that need to be estimated. w_k is the weight coefficient and is usually modeled using the Beta function to ensure that the sum of the weights is 1. These weights determine the impact of different lagging macroeconomic variables on the current long-term volatility. X_{t-k} is a lagging macroeconomic variable that represents a macroeconomic indicator for the $t-k$ period.

In order to ensure the non-negativity and normalization of the weight w_k, the Beta weight function is usually used for modeling, which is expressed as Equation 3.

$$w_k = \frac{(\frac{k}{K})^{\alpha_1-1}(\frac{1-k}{K})^{\alpha_2-1}}{\sum_{j=0}^{K}(\frac{j}{K})^{\alpha_1-1}(\frac{1-j}{K})^{\alpha_2-1}} \tag{3}$$

where α_1 and α_2 are the shape parameters of the Beta distribution, which need to be determined by estimation.

The advantage of the MIDAS regression model is that it can synthesize data of different frequencies in one model, so as to make full use of the information of macroeconomic variables and capture the dynamic changes in long-term volatility. The total conditional variance is expressed as Equation 4.

$$\sigma_t^2 = \tau_t \times g_t \tag{4}$$

Where σ_t^2 is the total conditional variance. τ_t is the long-term volatility component and is modeled using MIDAS regression. g_t is the short-term volatility component and follows the GARCH process.

Combining the advantages of GARCH and MIDAS, we get the GARCH-MIDAS model. The core idea is to treat short-term volatility and long-term volatility separately and model them separately, thereby improving the accuracy and reliability of volatility forecasting.

4. Experiments

4.1. Experimental Setups

Using the S&P 500 daily return and a series of macroeconomic variables, this study uses the GARCH-MIDAS model to analyze the relationship between macroeconomic indicators and stock market volatility. We first use the GARCH(1,1) model to estimate the short-term volatility, then correlate the long-term volatility with low-frequency macroeconomic variables through MIDAS regression, and finally integrate the short-term and long-term volatility components to construct a total conditional variance model. Through both intra- and out-of-sample data validation, we evaluated the prediction accuracy and robustness of the model. Figure 1 shows the used dataset.

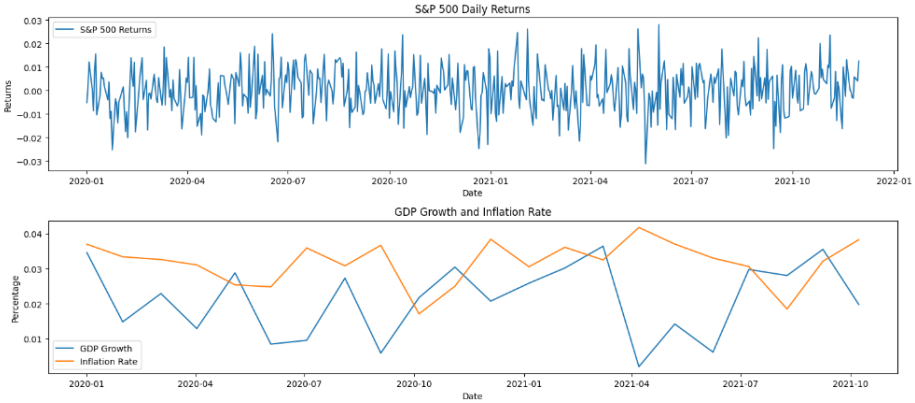

Figure 1. Illustration of Used Dataset.

DOI: 10.54254/2755-2721/87/20241611

4.2. Experimental Analysis

Volatility forecast error is used to assess the difference between the predicted volatility and the actual volatility. This metric measures the accuracy of a model in predicting market volatility. The smaller volatility prediction error indicates that the model is able to capture the volatility of the actual market more accurately, providing more reliable risk assessment and decision support. By calculating the error between the predicted volatility and the actual volatility, we can judge the prediction effect of the model and make corresponding improvements and optimizations. Figure 2 shows the volatility forecast error comparison results.

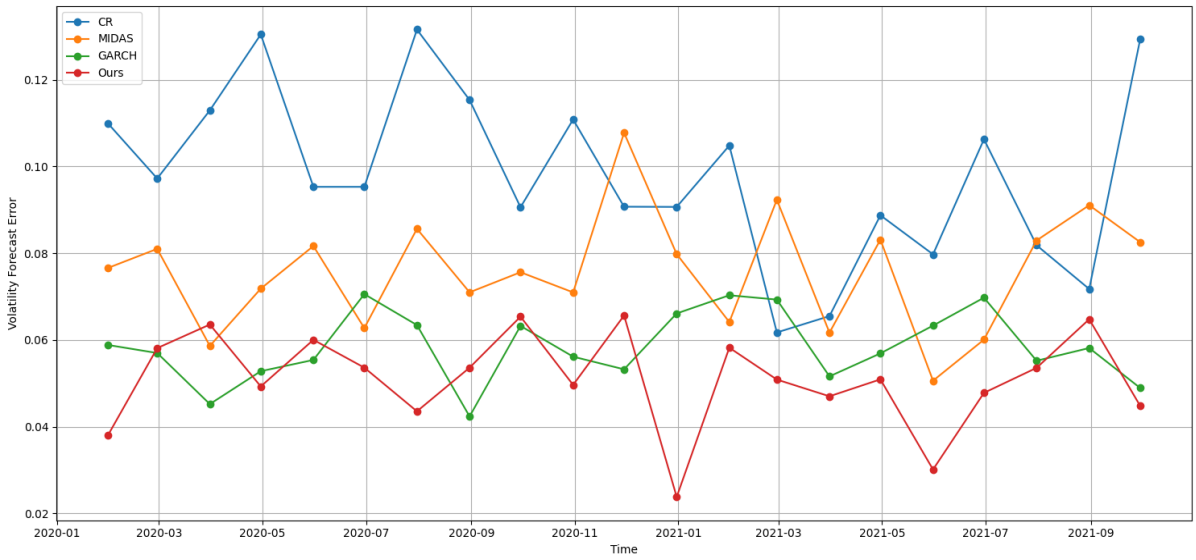

Figure 2. Volatility Forecast Error Comparison.

The confusion matrix is a tool used to evaluate the performance of a classification model, which shows in detail the performance of the model on each classification by comparing the predicted results with the actual results. The confusion matrix contains four key metrics: true positives, false positives, true negatives, and false negatives. Figure 3 shows the confusion matrix of our proposed model.

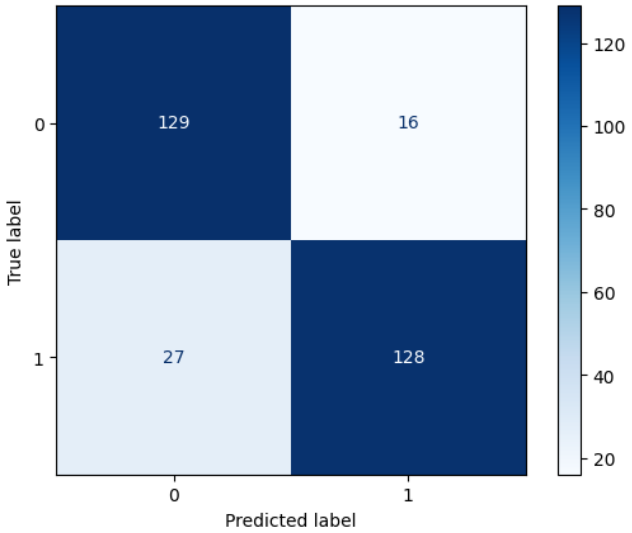

Figure 3. Confusion Matrix.

DOI: 10.54254/2755-2721/87/20241611

5. Conclusion

In conclusion, the GARCH-MIDAS model, an advanced machine learning technology, is used to analyze the relationship between macroeconomic indicators and stock market volatility, which significantly improves the ability to capture short-term and long-term volatility dynamics. By effectively integrating high-frequency equity market data and low-frequency macroeconomic variables, our approach provides more accurate and reliable volatility forecasts. Experiments verify the superior performance of our method compared with the traditional model. This comprehensive analysis provides valuable insights for investors, policymakers, and economists, enhancing risk management and decision-making processes in financial markets.

References

[1] Khan, Muhammad Kamran, et al. "Stock market reaction to macroeconomic variables: An assessment with dynamic autoregressive distributed lag simulations." International Journal of Finance & Economics 28.3 (2023): 2436-2448.

[2] Chang, Bisharat Hussain, et al. "Macroeconomic variables and stock indices: an asymmetric evidence from quantile ARDL model." South Asian Journal of Business Studies 10.2 (2021): 242-264.

[3] Hashmi, Shabir Mohsin, and Bisharat Hussain Chang. "Asymmetric effect of macroeconomic variables on the emerging stock indices: A quantile ARDL approach." International Journal of Finance & Economics 28.1 (2023): 1006-1024.

[4] Boukhatem, Jamel, Zied Ftiti, and Jean Michel Sahut. "Bond market and macroeconomic stability in East Asia: a nonlinear causality analysis." Annals of Operations Research 297.1 (2021): 53-76.

[5] Sun, Hongxiang, Zhongkai Yao, and Qingchun Miao. "Design of macroeconomic growth prediction algorithm based on data mining." Mobile Information Systems 2021.1 (2021): 2472373.

[6] Ma, Yaming, Ziwei Wang, and Feng He. "How do economic policy uncertainties affect stock market volatility? Evidence from G7 countries." International Journal of Finance & Economics 27.2 (2022): 2303-2325.

[7] Nagao, Ryoya, Yoshihiro Kondo, and Yoshiyuki Nakazono. "The macroeconomic effects of monetary policy: Evidence from Japan." Journal of the Japanese and International Economies 61 (2021): 101149.

[8] Bhargava, Vivek, and Daniel Konku. "Impact of exchange rate fluctuations on US stock market returns." Managerial finance 49.10 (2023): 1535-1557.

[9] Li, Yun Daisy, Talan B. İşcan, and Kuan Xu. "The impact of monetary policy shocks on stock prices: Evidence from Canada and the United States." Journal of international money and finance 29.5 (2010): 876-896.

[10] Kim, Yunmi, and Charles R. Nelson. "Pricing stock market volatility: does it matter whether the volatility is related to the business cycle?." Journal of Financial Econometrics 12.2 (2013): 307-328.

[11] Ghysels, Eric, Pedro Santa-Clara, and Rossen Valkanov. "The MIDAS touch: Mixed data sampling regression models." (2004).

[12] Engle, Robert F., Eric Ghysels, and Bumjean Sohn. "Stock market volatility and macroeconomic fundamentals." Review of Economics and Statistics 95.3 (2013): 776-797.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241591

Sound feature analysis and gender recognition based on deep learning: A review

Chenxu Zhu

School of, University of Liverpool, UK

sgczhu10@liverpool.ac.uk

Abstract. The application of deep learning in identifying sound features has become increasingly prevalent, greatly enhancing the performance of voice applications across various professional domains. This study focuses on deep learning techniques applied to sound feature analysis and gender recognition. It reviews methodologies, datasets, and case studies, emphasizing deep learning's crucial role in boosting efficiency and accuracy. Recent advancements highlight CNN-based architectures and novel models, demonstrating deep learning for voice systems for enhanced interaction and analysis. Challenges such as computational demands and limited data availability persist, but ongoing optimizations and multi-modal approaches promise future advancements in voice technology, enabling more intelligent and responsive interactions.

Keywords: Deep learning, Sound feature analysis, Gender recognition, Mel-Frequency Cepstral Coefficients (MFCC), Voice systems.

1. Introduction

The application of deep learning in identifying sound features has emerged as a mainstream trend, significantly enhancing the performance of various voice applications and user experiences across diverse professional fields [1] [2] [3] [4]. This advancement enables the processing and recognition of extensive sound data within shorter timeframes, thereby substantially improving the accuracy of related tasks.

Sound characteristics are primarily determined by the time domain, frequency domain, and time-frequency domain. Traditional methods, such as Mel-Frequency Cepstral Coefficients (MFCC), often face limitations in recognition effectiveness and accuracy [5] [6]. In contrast, deep learning can process larger datasets and optimize associated algorithms and outcomes, rendering it indispensable for tasks such as voice type and gender recognition. Deep learning excels in automatically learning complex features from data through the construction and training of deep neural networks.

Recent studies have extensively reviewed deep learning for sound classification. Some researchers critically evaluated recent research advancements concerning small data, specifically focusing on the use of data augmentation methods to increase the data available for deep learning classifiers in sound classification, including voice, speech, and related audio signals [2]. Besides, others presented a state-of-the-art review of various convolutional neural network (CNN) approaches in the audio domain, identifying challenges for sound classification systems [3]. Some experts also investigated the architecture and applications of deep learning in audio classification, providing an extensive review of existing research on audio-based techniques and discussing current limitations while proposing

DOI: 10.54254/2755-2721/87/20241591

directions for future research in audio-based deep learning methods [4]. These reviews consistently demonstrate the efficiency and accuracy of deep learning in sound classification. CNN and Recurrent Neural Networks (RNN) are two deep learning methods frequently employed to capture intricate patterns and time-dependent relationships within sound signals. These methods streamline the cumbersome processes of feature extraction and classification inherent in traditional methods, making them highly effective for sound classification tasks.

This review aims to identify the gender of the voice owner utilizing deep learning techniques, encompassing current development prospects, challenges, and future research directions. By examining the methodologies, datasets, and practical case studies, this review seeks to illustrate the pivotal role of deep learning in sound feature recognition. The focus will be on enhancing the efficiency and accuracy of voice recognition, improving the security and completeness of fields employing this technology, and elevating the overall quality of service.

2. Literature Survey

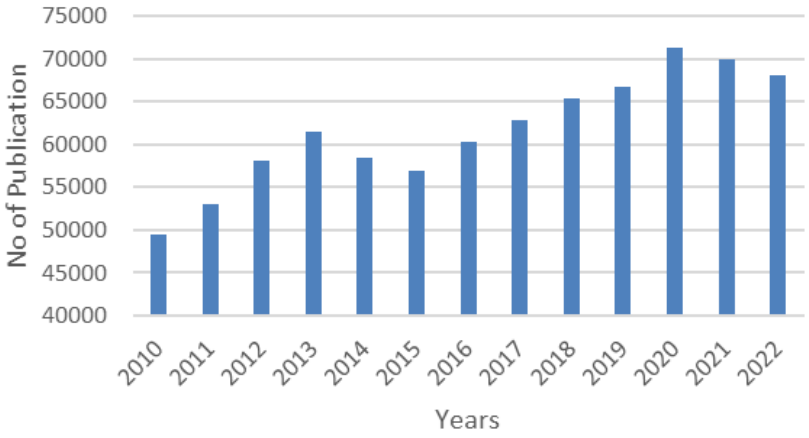

Figure 1. The number of papers searched using “deep learning” and “sound feature” per year.

Figure 1 illustrates the total number of papers retrieved using the search terms "deep learning" and "sound features" on Google Scholar from 2010 to 2022. The data reveals a gradual increase in the number of publications over this period, indicating growing research interest in this field. The number of papers rose from 49,500 in 2010 to a peak of 71,200 in 2020. This trend demonstrates that there has been consistent and increasing interest in the application of deep learning to sound features across the years.

3. Sound Features for Enhanced Gender Identification

Sound is made up of many features, each with different effects on audio processing and analysis. Mel-Frequency Cepstral Coefficients (MFCCs) are a commonly used feature in speech and audio processing, representing a scale of pitch that is perceived by the listener as aurally equal. MFCCs capture the power spectrum of a signal in a way that approximates the critical band structure of the human ear, making them highly effective in speech and speaker recognition tasks [7] [8].

Pitch, also known as the Fundamental Frequency, is another commonly used sound feature. It is a perceptual correlate of the fundamental frequency of a speech signal, conveying important information about the speaker such as intonation, stress, and emotional state. Pitch is often used in gender identification due to the different pitch ranges of men and women [9] [10].

Formants are the resonant frequencies of the vocal tract, and they differ between men and women due to variations in the vocal tract's physical structure [11] [12]. Other features, such as Spectral Features,

DOI: 10.54254/2755-2721/87/20241591

Zero-Crossing Rate (ZCR), and Linear Predictive Coding (LPC) Coefficients, also play significant roles in gender identification [13].

The process of computing MFCCs involves several steps: pre-emphasis, framing, windowing, Fast Fourier Transform (FFT), Mel filter bank processing, and Discrete Cosine Transform (DCT). Researchers can optimize MFCC performance by adjusting the number of Mel filter banks or selecting specific DCT coefficients to enhance the accuracy of gender identification [14].

Regarding Pitch and Formants, research can analyze the differences in the fundamental frequency range and formant frequencies between speakers of different genders. Spectral features, such as spectral centroid, bandwidth, and flatness, provide information about the shape of the sound spectrum. Comparing these spectral features for distribution differences is also a part of the processing. Additionally, the variation of ZCR in different speech segments, such as vowels and consonants, is an important indicator for evaluation.

4. Gender Recognition by Deep Learning

Deep learning plays a crucial role in sound feature analysis owing to its capability in processing complex data and extracting high-dimensional features. Within this domain, CNNs and RNNs are prominent architectures used for voice feature analysis and gender recognition.

CNNs, typically employed in image processing, are adept at analyzing sound features such as spectral graphs and MFCCs. Through convolutional layers, CNNs extract local features, while pooling layers effectively reduce the dimensionality of feature maps to capture essential time-frequency information [15], thereby achieving high-precision gender identification. On the other hand, RNNs excel in processing sequential data. In sound feature analysis, RNNs capture temporal sequences in speech signals, such as dynamic changes in fundamental frequency tracks, thereby enhancing gender recognition accuracy [16].

Preprocessing sound features is pivotal in deep learning pipelines as it enhances feature quality, consequently improving model accuracy and robustness. Key preprocessing techniques include:

1. Signal Denoising: Utilizing filtering and adaptive noise cancellation methods to enhance feature extraction quality and speech signal clarity.
2. Pre-emphasis: Applying a pre-emphasis filter to boost high-frequency components of voice signals [17].
3. Frame Segmentation and Windowing: Dividing speech signals into short-time frames to capture local time-frequency characteristics effectively.
4. Fourier Transform: Employing Fast Fourier Transform (FFT) to convert time-domain signals into frequency-domain representations, facilitating spectral analysis of each frame [18].
5. MFCC Extraction: Processing the spectrum through a Mel filter bank, computing the logarithmic energy output of filters, and applying Discrete Cosine Transform (DCT) to obtain MFCCs.
6. Spectrogram Generation: Using Short-Time Fourier Transform (STFT) to generate spectrograms of speech signals, facilitating CNN processing.
7. Normalization: Adjusting feature values to a consistent range to mitigate variations during model training, thereby enhancing convergence and performance [19] [20].

5. Advancements in Deep Learning for Voice Systems

AI-driven voice systems perform diverse tasks by accurately interpreting user commands, benefiting from deep learning's efficiency in feature extraction and command recognition. CNNs extract high-dimensional features, RNNs process time series features, and Transformer models with self-attention mechanisms efficiently recognize user commands and classify datasets derived from spectrograms. Self-attention mechanisms and Graph Convolutional Networks (GCNs) further enhance performance in speech recognition, music classification, and speaker recognition.

Recent research highlights various aspects of deep learning in voice systems. The researchers discussed the effectiveness of AI systems in handling simple and complex service requests, noting reduced customer complaints with prior user experience [21]. Some laboratories developed an intelligent

DOI: 10.54254/2755-2721/87/20241591

wheelchair controlled by CNN-based voice commands, aiding mobility for individuals with disabilities [22]. Besides, the expert introduced the Time-Frequency Capsule Neural Network (TFCap) for stable global information extraction from spectrograms, evaluated on the IEMOCAP database [23]. The researchers also reviewed convolutional feature extraction methods for deep neural network-based sound source localization [24]. In addition, a CNN-based approach integrating pre-processing, feature extraction, reduction, and classification stages for sound analysis had been processed by some scholars [25]. Another study demonstrated the enhancement of environmental sound classification using a convolutional RNN combined with a frame-level attention mechanism for discriminative feature learning [26]. These studies illustrate the broad application of deep learning in voice systems, encompassing device control, localization tasks, innovative methodologies, and refined environmental sound analysis.

6. Challenges and Future Directions

Deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are complex and parameter-heavy, demanding significant computational resources and time for training [27] [28]. This poses challenges for researchers and institutions with limited computational capabilities. Moreover, these models require large volumes of labeled data, which can be particularly challenging in voice feature analysis and gender identification due to data scarcity and variations in data distribution across different application scenarios, hindering model generalization.

Despite these challenges, ongoing advancements in algorithm optimization and computing power are expected to enhance the accuracy and efficiency of deep learning models in voice feature analysis and gender recognition. Future developments may introduce more sophisticated neural network architectures capable of handling increasingly complex data, thereby pushing the field forward. Additionally, integrating multi-modal data (e.g., audio, text, and image) into deep learning models is poised to expand research horizons, aiming for more diverse and versatile applications [29].

The integration of voice feature analysis and gender recognition technologies with existing speech recognition systems holds promise for advancing voice applications [30]. Deep learning in feature analysis and the real-time processing strengths of speech recognition technology, future systems can deliver more intelligent and responsive voice interactions.

7. Summary

This review explores the pivotal role of sound features in gender identification. Recent advancements highlight CNN-based devices and innovative models, showcasing the integration of deep learning in voice systems for improved interaction and analysis. Challenges include computational demands and data scarcity, but ongoing optimizations and multi-modal approaches promise future advancements in voice technology and application integration, ensuring more intelligent and responsive voice interactions.

References

[1] Hu, H. C., Chang, S. Y., Wang, C. H., Li, K. J., Cho, H. Y., Chen, Y. T., ... & Lee, O. K. S. (2021). Deep learning application for vocal fold disease prediction through voice recognition: preliminary development study. Journal of medical Internet research, 23(6), e25247.

[2] Abayomi-Alli, O. O., Damaševičius, R., Qazi, A., Adedoyin-Olowe, M., & Misra, S. (2022). Data augmentation and deep learning methods in sound classification: A systematic review. Electronics, 11(22), 3795.

[3] Bhattacharya, S., Das, N., Sahu, S., Mondal, A., & Borah, S. (2021). Deep classification of sound: A concise review. In Proceeding of First Doctoral Symposium on Natural Computing Research: DSNCR 2020 (pp. 33-43). Springer Singapore.

[4] Wang, Y., Wei-Kocsis, J., Springer, J. A., & Matson, E. T. (2022, October). Deep learning in audio classification. In International Conference on Information and Software Technologies (pp. 64-77). Cham: Springer International Publishing.

[5] Tiwari, V. (2010). MFCC and its applications in speaker recognition. International journal on. emerging technologies, 1(1), 19-22.
[6] Mohammed, R. A., Ali, A. E., & Hassan, N. F. (2019). Advantages and disadvantages of automatic speaker recognition systems. Journal of Al-Qadisiyah for computer science and mathematics, 11(3), Page-21.
[7] Abdul, Z. K., & Al-Talabani, A. K. (2022). Mel frequency cepstral coefficient and its applications: A review. IEEE Access, 10, 122136-122158.
[8] Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. arXiv preprint arXiv:1003.4083.
[9] Sharma, G., Umapathy, K., & Krishnan, S. (2020). Trends in audio signal feature extraction methods. Applied Acoustics, 158, 107020.
[10] Hirst, D. J., & de Looze, C. (2021). Measuring Speech. Fundamental frequency and pitch. Cambridge Handbook of Phonetics, (1), 336-361.
[11] Aalto, D., Malinen, J., & Vainio, M. (2018). Formants. In Oxford Research Encyclopedia of Linguistics.
[12] Schafer, R. W., & Rabiner, L. R. (1970). System for automatic formant analysis of voiced speech. The Journal of the Acoustical Society of America, 47(2B), 634-648.
[13] Chauhan, N., Isshiki, T., & Li, D. (2019, February). Speaker recognition using LPC, MFCC, ZCR features with ANN and SVM classifier for large input database. In 2019 IEEE 4th international conference on computer and communication systems (ICCCS) (pp. 130-133). IEEE.
[14] Sidhu, M. S., Latib, N. A. A., & Sidhu, K. K. (2024). MFCC in audio signal processing for voice disorder: a review. Multimedia Tools and Applications, 1-21.
[15] O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.
[16] Son, G., Kwon, S., & Park, N. (2019). Gender classification based on the non-lexical cues of emergency calls with recurrent neural networks (RNN). Symmetry, 11(4), 525.
[17] Schnell, K., & Lacroix, A. (2007, August). Time-varying pre-emphasis and inverse filtering of speech. In INTERSPEECH (pp. 530-533).
[18] Heckbert, P. (1995). Fourier transforms and the fast Fourier transform (FFT) algorithm. Computer Graphics, 2(1995), 15-463.
[19] Serbes, G., Ulukaya, S., & Kahya, Y. P. (2018). An automated lung sound preprocessing and classification system based onspectral analysis methods. In Precision Medicine Powered by pHealth and Connected Health: ICBHI 2017, Thessaloniki, Greece, 18-21 November 2017 (pp. 45-49). Springer Singapore.
[20] Oh, W. (2020). Comparison of environmental sound classification performance of convolutional neural networks according to audio preprocessing methods. The Journal of the Acoustical Society of Korea, 39(3), 143-149.
[21] Wang, L., Huang, N., Hong, Y., Liu, L., Guo, X., & Chen, G. (2023). Voice-based AI in call center customer service: A natural field experiment. Production and Operations Management, 32(4), 1002-1018.
[22] Sharifuddin, M. S. I., Nordin, S., & Ali, A. M. (2019, September). Voice control intelligent wheelchair movement using CNNs. In 2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS) (pp. 40-43). IEEE.
[23] Liu, J., Song, Y., Wang, L., Dang, J., & Yu, R. (2021). Time-Frequency Representation Learning with Graph Convolutional Network for Dialogue-Level Speech Emotion Recognition. In Interspeech (pp. 4523-4527).
[24] Krause, D., Politis, A., & Kowalczyk, K. (2021, January). Comparison of convolution types in CNN-based feature extraction for sound source localization. In 2020 28th European Signal Processing Conference (EUSIPCO) (pp. 820-824). IEEE.

DOI: 10.54254/2755-2721/87/20241591

[25] Demir, F., Turkoglu, M., Aslan, M., & Sengur, A. (2020). A new pyramidal concatenated CNN approach for environmental sound classification. Applied Acoustics, 170, 107520.
[26] Zhang, Z., Xu, S., Zhang, S., Qiao, T., & Cao, S. (2021). Attention based convolutional recurrent neural network for environmental sound classification. Neurocomputing, 453, 896-903.
[27] Frikha, M., Taouil, K., Fakhfakh, A., & Derbel, F. (2022). Limitation of deep-learning algorithm for prediction of power consumption. Engineering Proceedings, 18(1), 26.
[28] Islam, M. S., Sultana, S., Kumar Roy, U., & Al Mahmud, J. (2020). A review on video classification with methods, findings, performance, challenges, limitations and future work. Jurnal Ilmiah Teknik Elektro Komputer dan Informatika, 6(2), 47-57.
[29] Amal, S., Safarnejad, L., Omiye, J. A., Ghanzouri, I., Cabot, J. H., & Ross, E. G. (2022). Use of multi-modal data and machine learning to improve cardiovascular disease care. Frontiers in cardiovascular medicine, 9, 840262.
[30] Deng, L. (2016). Deep learning: from speech recognition to language and multimodal processing. APSIPA Transactions on Signal and Information Processing, 5, e1.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241603

Application of artificial intelligence in cancer imaging diagnosis: A review

Daochun Chen

School of Mechanical and Electronic Information, China University of Geosciences (Wuhan), 388 Lumo Road, Wuhan, Hubei, 430074, China

1983567421@qq.com

Abstract. Medical imaging analysis is integral to modern clinical practice, providing crucial insights into internal anatomical structures essential for disease diagnosis and treatment planning. Traditional diagnostic methods often rely on subjective interpretation, leading to inconsistencies and delays. In recent years, artificial intelligence (AI) has revolutionized medical imaging by enhancing diagnostic accuracy and efficiency. AI technologies, particularly deep learning algorithms, process vast datasets to uncover patterns and anomalies, improving lesion detection and classification. This review explores the application of AI on cancer imaging diagnosis, highlighting advancements in image analysis, including lesion detection, segmentation, and feature extraction. It examines the integration of AI with omics technologies for comprehensive patient profiling and personalized treatment strategies. Moreover, the review discusses future directions and ethical considerations, underscoring AI's potential to reshape cancer diagnosis and improve patient outcomes.

Keywords: Artificial intelligence, Cancer, Image diagnosis.

1. Introduction

Medical imaging analysis plays a pivotal role in clinical diagnosis and treatment by providing detailed insights into the body's internal structures, essential for accurately identifying diseases. Traditional diagnostic approaches often rely on expertise and professional judgment from doctors, utilizing symptom observation, clinical signs, and lab results. However, the subjectivity inherent in this method can lead to inconsistencies and inaccuracies due to personal bias and varying experiences. Moreover, manually analyzing medical records and test findings is time-intensive, potentially causing delays in diagnosis, especially when handling a large volume of cases.

In recent years, the integration of artificial intelligence (AI) into medical image analysis has significantly improved lesion detection and diagnostic accuracy [1, 2, 3]. Innovations such as image enhancement and multimodal image fusion have further elevated medical imaging, providing comprehensive and precise diagnostic information. AI technology leverages machine learning and deep learning algorithms to process medical data, offering more precise diagnoses. By analyzing extensive datasets, AI can uncover patterns and trends, enhancing diagnostic accuracy and efficiency. With its automation capabilities, AI can swiftly interpret medical images and test results, reducing the workload on doctors and boosting productivity. AI technologies, particularly deep learning, offer powerful tools for efficiently managing and interpreting large quantities of imaging data [4, 5, 6]. They enhance

DOI: 10.54254/2755-2721/87/20241603

diagnostic precision and speed, reduce human error, and enable the early detection of conditions such as cancer, which is crucial for improving patient outcomes and survival rates.

In cancer imaging, AI systems analyze image data to detect anomalies, classify different tumor types, and predict disease progression. Deep learning employs neural networks that mimic the structure and function of the human brain, making them particularly adept at processing high-dimensional data such as detailed tumor images. Through tasks like image segmentation, lesion detection, and feature extraction, AI accurately identifies cancerous regions, aiding doctors in making informed decisions [7, 8, 9]. The application of AI in cancer imaging diagnosis has been extensively studied. Bi reviewed the current state of AI in medical imaging for cancer, describing advancements in lung, brain, breast, and prostate tumors to illustrate how common clinical challenges are being addressed [10]. Sadoughi examined various AI techniques utilizing medical images for breast cancer detection [11]. Hunter discussed how AI algorithms assist clinicians in screening asymptomatic patients at risk of cancer, investigating and triaging symptomatic patients, and more effectively diagnosing cancer recurrence [12]. Barragán-Montero presented the foundational technological pillars of AI and state-of-the-art machine learning methods, discussing new trends and future research directions in medical imaging [13]. Koh outlined relevant AI and machine learning techniques, highlighting key opportunities for implementing these technologies in cancer imaging [14]. Their research demonstrates the diverse applications of AI in cancer imaging diagnosis, exploring how AI can improve the accuracy, efficiency, and clinical practice of cancer diagnosis, and firmly confirms the potential for further application and development of AI technology in this field.

This review explores the intersection of artificial intelligence technology and cancer imaging diagnosis, aiming to comprehensively outline the current progress, challenges, solutions, and future directions in this domain. By analyzing existing research findings and clinical practices, the review seeks to offer theoretical and practical support for furthering the application of AI technology in cancer imaging diagnosis. Ultimately, the goal is to advance the detection and treatment of early-stage cancer, thereby enhancing survival rates and the quality of life for patients.

2. Literature Survey

Figure 1 illustrates the annual number of papers retrieved using the search terms “Artificial Intelligence” and “Cancer Imaging Diagnosis” on Google Scholar. The overall trend of relevant literature has been on the rise from 2010 to 2023. The number of papers rose from 10200 in 2010 to a peak of 56400 in 2022. It should be noted that although the number of papers in 2023 is slightly lower than that in 2022, the difference is almost the same. This trend indicates a growing interest in the application of Artificial Intelligence in Cancer Imaging Diagnosis.

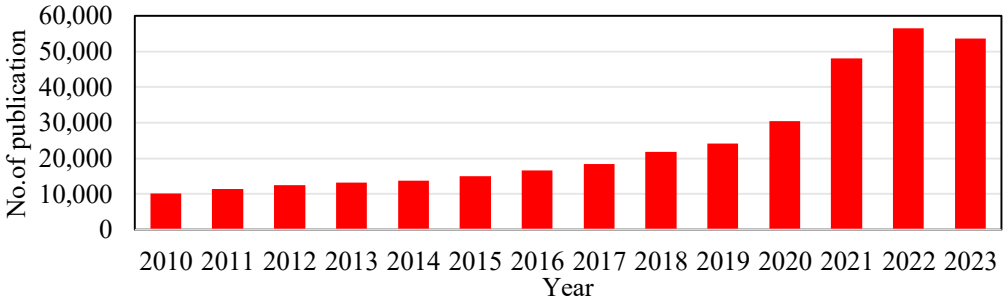

Figure 1. The number of papers searched using “Artificial Intelligence” and “Cancer Imaging Diagnosis” per year

DOI: 10.54254/2755-2721/87/20241603

3. Advanced Applications of AI in Medical Imaging for Cancer Diagnosis

3.1. Medical image analysis and omics

AI technology plays a pivotal role in analyzing diverse medical images such as X-rays, CT scans, and MRIs for early cancer detection, lesion analysis, classification, and treatment monitoring. Integration with omics technology enables rapid analysis of comprehensive patient data, facilitating precise diagnoses and tailored treatment plans. Tumors can be categorized based on characteristics like size and spread, providing crucial insights for clinical decision-making. AI-powered techniques in cancer imaging accurately assess tumor size, shape, texture, and dynamics [15, 16].

Zheng explored AI-driven imaging techniques including mammography, ultrasound, MRI, and PET for breast cancer screening and diagnosis [17]. Telecan discussed advanced decision support tools utilizing texture analysis and AI for MRI image analysis, aiding in prostate cancer diagnosis, staging, aggressiveness prediction, and biopsy guidance [18]. Dynamic contrast uptake assessment in MRI facilitates the characterization of tumor masses by heterogeneity, spatial phenotype, and dynamic features. Systems biology algorithms, combining AI with omics technology, expedite and enhance the accuracy of patient data analysis, supporting precise diagnoses and tailored treatment plans.

3.2. Computer auxiliary diagnosis system

Computer-Aided Diagnosis (CAD) systems are indispensable tools for radiologists in cancer imaging, offering functions such as automatic lesion detection and diagnostic decision support through AI technology. This technological advancement facilitates early cancer detection, reduces the risk of missed or incorrect diagnoses, and advances medical imaging towards intelligent and precise diagnostics [19, 20].

Yao addressed limitations of traditional AI models in breast cancer diagnosis through machine learning, enhancing early detection accuracy and reducing misdiagnosis rates by providing clear medical images and computer-aided diagnosis [21]. Arun and Sasikala focused on deep learning techniques to improve breast cancer detection, exploring architectures like CNN, transfer learning, cross-modal learning, and fine-tuning CNN [22]. Optimization of hyperparameters and effective feature selection strategies can enhance the performance of CAD systems.

3.3. Image segmentation and feature extraction

Deep learning algorithms are instrumental in automating tumor image segmentation and feature extraction, thereby enhancing diagnostic accuracy for healthcare providers. These algorithms efficiently identify patterns and features in medical images, enabling clearer observation of tumor morphology, size, and texture, critical for precise diagnosis and treatment [23, 24].

Tang compared various methodologies for AI nodule segmentation, feature extraction, and classification, including determining optimal deep learning segmentation techniques for lung nodules, employing the Image Biomarker Standardization Initiative (IBSI) for feature extraction, and utilizing principal component analysis (PCA) alongside different machine learning techniques to identify optimal methodologies based on extracted features [25]. Ranjbarzadeh reviewed AI applications in brain tumor segmentation, demonstrating proficiency in distinguishing between abnormal and normal brain tissue [26]. These technologies showcase precision and utility in healthcare settings, particularly in enhancing diagnostic workflows.

4. Future Directions

AI technology is advancing cancer diagnosis by integrating with fields like genomics and molecular diagnostics, expanding its role in medical imaging for precise identification, classification, and personalized treatment planning across various cancer types. This integration incorporates genomic data

DOI: 10.54254/2755-2721/87/20241603

to enhance understanding of patient genetic profiles, assess risks, and customize treatment strategies. By combining biomarkers and molecular signals with imaging technology, AI improves accuracy in cancer classification, grading, and prognosis. This holistic approach revolutionizes cancer imaging diagnosis, providing precise tools for personalized care and advancing medical diagnostics. Looking forward, AI promises continued advancements in diagnostic precision and efficiency, poised to transform healthcare with enhanced image data mining and analysis capabilities. While offering substantial benefits, the adoption of AI raises ethical considerations such as safeguarding patient privacy, ensuring transparency and interpretability in AI decision-making, and addressing potential biases. It is essential to provide comprehensive ethical training to healthcare professionals and technicians to guide the responsible integration of AI in medical imaging, ensuring adherence to ethical standards and maximizing benefits for patients and society.

5. Summary

Medical imaging analysis is crucial for clinical diagnosis and treatment, offering detailed insights into internal bodily structures essential for identifying diseases accurately. Traditional diagnostic methods rely heavily on clinical expertise and subjective judgment, leading to potential inconsistencies and delays. In recent years, AI has revolutionized medical imaging by enhancing lesion detection and diagnostic precision. AI technologies, particularly deep learning, analyze extensive datasets to uncover patterns, improving efficiency and reducing errors. In cancer imaging, AI aids in anomaly detection, tumor classification, and disease progression prediction, significantly advancing early diagnosis and treatment efficacy. Integrating AI with omics technologies further enhances diagnostic capabilities, offering tailored treatment plans based on comprehensive patient data. As AI continues to evolve, it promises to transform cancer imaging by refining diagnostic accuracy and personalized care, revolutionizing healthcare outcomes.

References

[1] Rana, M., & Bhushan, M. (2023). Machine learning and deep learning approach for medical image analysis: diagnosis to detection. Multimedia Tools and Applications, 82(17), 26731-26769.

[2] Aleid, A., Alhussaini, K., Alanazi, R., Altwaimi, M., Altwijri, O., & Saad, A. S. (2023). Artificial intelligence approach for early detection of brain tumors using MRI images. Applied Sciences, 13(6), 3808.

[3] Hampiholi, N. (2023). Medical Imaging Enhancement with Ai Models for Automatic Disease Detection and Classification Based on Medical Images. International Journal of Engineering Applied Sciences and Technology, 8(5), 31-37.

[4] Li, R., Zhang, W., Suk, H. I., Wang, L., Li, J., Shen, D., & Ji, S. (2014). Deep learning based imaging data completion for improved brain disease diagnosis. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part III 17 (pp. 305-312). Springer International Publishing.

[5] Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., & Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. Journal of Medical Imaging and Radiation Oncology, 65(5), 545-563.

[6] Kebaili, A., Lapuyade-Lahorgue, J., & Ruan, S. (2023). Deep learning approaches for data augmentation in medical imaging: a review. Journal of Imaging, 9(4), 81.

[7] Jaleel, J. A., Salim, S., & Aswin, R. (2012). Artificial neural network based detection of skin cancer. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 1(3).

[8] Vijayalakshmi, M. M. (2019). Melanoma skin cancer detection using image processing and machine learning. International Journal of Trend in Scientific Research and Development (IJTSRD), 3(4), 780-784.

[9] Wang, W., Bu, F., Lin, Z., & Zhai, S. (2020). Learning methods of convolutional neural network combined with image feature extraction in brain tumor detection. IEEE access, 8, 152659-152668.

[10] Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., ... & Aerts, H. J. (2019). Artificial intelligence in cancer imaging: clinical challenges and applications. CA: a cancer journal for clinicians, 69(2), 127-157.

[11] Sadoughi, F., Kazemy, Z., Hamedan, F., Owji, L., Rahmanikatigari, M., & Azadboni, T. T. (2018). Artificial intelligence methods for the diagnosis of breast cancer by image processing: a review. Breast Cancer: Targets and Therapy, 219-230.

[12] Hunter, B., Hindocha, S., & Lee, R. W. (2022). The role of artificial intelligence in early cancer diagnosis. Cancers, 14(6), 1524.

[13] Barragán-Montero, A., Javaid, U., Valdés, G., Nguyen, D., Desbordes, P., Macq, B., ... & Lee, J. A. (2021). Artificial intelligence and machine learning for medical imaging: A technology review. Physica Medica, 83, 242-256.

[14] Koh, D. M., Papanikolaou, N., Bick, U., Illing, R., Kahn Jr, C. E., Kalpathi-Cramer, J., ... & Prior, F. (2022). Artificial intelligence and machine learning in cancer imaging. Communications Medicine, 2(1), 133.

[15] Fusco, R., Piccirillo, A., Sansone, M., Granata, V., Vallone, P., Barretta, M. L., ... & Petrillo, A. (2021). Radiomic and artificial intelligence analysis with textural metrics, morphological and dynamic perfusion features extracted by dynamic contrast-enhanced magnetic resonance imaging in the classification of breast lesions. Applied Sciences, 11(4), 1880.

[16] Lv, T., Hong, X., Liu, Y., Miao, K., Sun, H., Li, L., ... & Pan, X. (2024). AI-powered interpretable imaging phenotypes noninvasively characterize tumor microenvironment associated with diverse molecular signatures and survival in breast cancer. Computer Methods and Programs in Biomedicine, 243, 107857.

[17] Zheng, D., He, X., & Jing, J. (2023). Overview of artificial intelligence in breast cancer medical imaging. Journal of Clinical Medicine, 12(2), 419.

[18] Telecan, T., Andras, I., Crisan, N., Giurgiu, L., Cătа, E. D., Caraiani, C., ... & Lupsor-Platon, M. (2022). More than meets the eye: using textural analysis and artificial intelligence as decision support tools in prostate cancer diagnosis—a systematic review. Journal of Personalized Medicine, 12(6), 983.

[19] Stoitsis, J., Valavanis, I., Mougiakakou, S. G., Golemati, S., Nikita, A., & Nikita, K. S. (2006). Computer aided diagnosis based on medical image processing and artificial intelligence methods. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 569(2), 591-595.

[20] Park, G. E., Kang, B. J., Kim, S. H., & Lee, J. (2022). Retrospective review of missed cancer detection and its mammography findings with artificial-intelligence-based, computer-aided diagnosis. Diagnostics, 12(2), 387.

[21] Yao, J., Zou, Y., Du, S., Wu, H., & Yuan, B. (2023). Progress in the Application of Artificial Intelligence in Ultrasound Diagnosis of Breast Cancer. Frontiers in Computing and Intelligent Systems, 6(1), 56-59.

[22] Arun Kumar, S., & Sasikala, S. (2023). Review on deep learning-based CAD systems for breast cancer diagnosis. Technology in Cancer Research & Treatment, 22, 15330338231177977.

[23] Sharmin, S., Ahammad, T., Talukder, M. A., & Ghose, P. (2023). A hybrid dependable deep feature extraction and ensemble-based machine learning approach for breast cancer detection. IEEE Access.

[24] Devunooru, S., Alsadoon, A., Chandana, P. W. C., & Beg, A. (2021). Deep learning neural networks for medical image segmentation of brain tumours for diagnosis: a recent review and taxonomy. Journal of Ambient Intelligence and Humanized Computing, 12, 455-483.

DOI: 10.54254/2755-2721/87/20241603

[25] Tang, T. W., Lin, W. Y., Liang, J. D., & Li, K. M. (2023). Artificial intelligence aided diagnosis of pulmonary nodules segmentation and feature extraction. Clinical Radiology, 78(6), 437-443.

[26] Ranjbarzadeh, R., Caputo, A., Tirkolaee, E. B., Ghoushchi, S. J., & Bendechache, M. (2023). Brain tumor segmentation of MRI images: A comprehensive review on the application of artificial intelligence tools. Computers in biology and medicine, 152, 106405.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241564

Music recommendation systems in music information retrieval: Leveraging machine learning and data mining techniques

Yan Chen

Jiangxi University of Applied Science, Jiangxi, China

lingwadesu@gmail.com

Abstract. Music Information Retrieval (MIR) has become a pivotal area of research with the rise of digital music platforms, enabling personalized music recommendations to enhance user experience. This paper explores the integration of machine learning and data mining techniques in music recommendation systems. We discuss user-based and item-based collaborative filtering, matrix factorization methods like Singular Value Decomposition (SVD) and Alternating Least Squares (ALS), and content-based filtering that incorporates audio feature analysis, metadata, and lyrics analysis. Additionally, we delve into hybrid recommendation systems, combining collaborative and content-based approaches using advanced models such as neural networks and hybrid autoencoders. Our finding show that, hybrid systems provide the most accurate and personalize recommendations, albeit requiring significant computational resources. Practical applications from platforms likes Spotify and Pandora illustrate the effectiveness of these approaches in real-world settings.

Keywords: Music Information Retrieval, Music Recommendation Systems, Machine Learning, Data Mining, Collaborative Filtering.

1. Introduction

The explosion of digital music platforms has drastically changed how people discover and enjoy music. With millions of tracks available at their fingertips, users rely on recommendation systems to navigate this vast sea of content. Music Information Retrieval (MIR) plays an important role in developing these recommendation systems, leveraging sophisticated algorithms to analyze and interpret musical data. This paper aims to provide a comprehensive overview of the current state of music recommendation systems, focusing on the integration of the machine learning and the data mining techniques. Traditional methods of music discovery, such as manual searches and curated playlists, are no longer sufficient to meet the diverse and dynamic tastes of modern listeners. User-based collaborative filterings, one of the earliest approaches, recommends music based on the listening behaviors of similar users. While effective, this method faces scalability issues as the number of users grows, necessitating the use of advanced techniques like clustering and approximate nearest neighbors to maintain efficiency. Item-based collaborative filtering shifts the focus from users to items, recommending songs that are frequently listened to together. This approach is more scalable but it still struggles with the cold start problem for new items. Matrix factorization techniques like SVD and ALS decompose the user-item interaction

DOI: 10.54254/2755-2721/87/20241564

matrix into latent factors, capturing underlying relationships and providing more accurate recommendations. These methods require substantial computational power and careful tuning to avoid overfitting. Content-based filtering offers another dimension, analyzing audio features, metadata, and lyrics to recommend similar music based on a user's past preferences. This method excels in recommending new or less popular songs but is limited by the quality of feature extraction. Techniques such as Mel-Frequency Cepstral Coefficients (MFCCs) and deep learning-based feature extraction significantly enhance the accuracy of music recommendations. Hybrid recommendation systems, which combine collaborative and content-based methods, leverage the strengths of both approaches. Model-based hybrid systems, employing neural networks and ensemble learning, can identify complex patterns from diverse data sources, thereby providing highly personalized recommendations. However, these models demand extensive training data and significant computational resources [1]. This paper also explores practical applications of these techniques, with case studies from platforms like Spotify and Pandora showcasing their effectiveness in real-world scenarios. By examining these advanced methods, we aim to illuminate future directions for music recommendation systems and their potential to revolutionize user experiences.

2. Collaborative Filtering

2.1. User-Based Collaborative Filtering

User-based collaborative filtering is a prevalent technique in music recommendation systems. It is based on the idea that users with similar listening histories will likely enjoy the same music. The system measures similarities between users by analyzing their listening behaviors and then recommends music that similar users have appreciated. For instance, if User A and User B have a high similarity score, and User A has recently enjoyed a new album, the system will likely recommend this album to User B. The primary advantage of this approach is its simplicity and the ability to generates diverse recommendations. However, it suffers from scalability issues as the number of users increases, and it requires substantial computational resources to maintain and update the similarity matrix. To address these issues, techniques such as clustering and approximate nearest neighbors can be employed to reduce computational complexity and improving efficiency [2]. Table 1 provides an overview of the listened songs by different users, their similarity score with User A, and the recommended songs for User B based on the similarity calculations. This data illustrates how user-based collaborative filtering works by leveraging the listening histories of similar users to generate personalized music recommendations.

Table 1. User-Based Collaborative Filtering Data

User ID	Listened Songs	Similarity Scores with User A	Recommended Songs for User B
User A	['Song 1', 'Song 2', 'Song 3', 'Song 4']	1	['Song 4']
User B	['Song 2', 'Song 3', 'Song 5', 'Song 6']	0.75	['Song 1', 'Song 4']
User C	['Song 1', 'Song 4', 'Song 7', 'Song 8']	0.5	['Song 2', 'Song 3']
User D	['Song 3', 'Song 5', 'Song 6', 'Song 9']	0.25	['Song 1', 'Song 4']

2.2. Item-Based Collaborative Filtering

Item-based collaborative filtering overcomes some of the limitations inherent in user-based approaches by emphasizing the relationships between items rather than users. In this approach, the system suggests music that is similar to tracks the user has previously enjoyed. This is accomplished by calculating the similarity between various songs or albums based on users' listening habits. For example, if a user often listens to both Song A and Song B, the system will recommend Song B to other users who have listened

DOI: 10.54254/2755-2721/87/20241564

to Song A [3]. This method tends to be more scalable and better suited for handling large datasets compared to user-based filtering. Nonetheless, it can still encounter the "cold start" problem when dealing with new items that lack sufficient interaction data. To address this issue, advanced techniques such as content-based filtering or hybrid methods can be employed, integrating additional information about the items to improve recommendations.

2.3. Matrix Factorization Techniques

Matrix factorization technique such as Singular Value Decomposition (SVD) and Alternating Least Squares (ALS), are advanced methods used to enhance collaborative filtering. These techniques decompose the user-item interaction matrix into latent factors, capturing the underlying relationships between users and items. By representing users and items in a lower-dimensional space, matrix factorization can uncover hidden pattern and provide more accurate recommendations. For example, SVD can predict a user's preference for a new song by combining their latent factors with those of the song. These techniques offer improve accuracy and scalability but require significant computational power and can be complex to implement and tune. Regularization techniques are often employed to prevent overfitting and improve the generalizations of the model. The formula for matrix factorization, specifically using Singular Value Decomposition (SVD), can be expressed as:

$$R \approx U \sum V_T \tag{1}$$

This formula decomposes the user-item interaction matrix R into three matrices—U, Σ, and V_T—capturing the latent factors that represent underlying relationships between users and items. By leveraging these latent factors, the recommendation system can predict a user's preferences for items with improved accuracy and scalability [4].

3. Content-Based Filtering

3.1. Audio Feature Analysis

Content-based filtering recommends music by analyzing the intrinsic audio features of songs, such as melody, rhythm, tempo, instrumentation, harmony, and timbre. This method focuses on the characteristics of the music itself rather than user interactions or preferences, enabling the system to create a detailed profile for each user based on the types of music they have previously enjoyed. For example, a system might analyze the spectral properties, such as Mel-Frequency Cepstral Coefficients (MFCCs), to capture the timbral texture of the music, or it might assess rhythmic patterns and tempo to understand the energetic qualities of the tracks. When a new song is introduced, its audio features are meticulously compared with the user's established profile to determine its relevance. For instance, if a user frequently listens to songs characterized by a high tempo and prominent guitar riffs, the recommendation system will prioritize new songs with similar audio features [5]. This approach is particularly advantageou for recommending new or less popular songs that might not have sufficient user interaction data to be included in collaborating filtering models. The effectiveness of content-based filtering lies in its ability to operate independently of user interactions, making it a robust tool for discovering emerging artists or niche genres that have not yet gained widespread popularity. However, the success of this approach is heavily dependent on the quality and granularity of the feature extraction process. Poorly extracted features can lead to inaccurate recommendations, while high-quality, detailed features can significantly enhances the system's performance. Advanced audio analysis techniques is crucial for improving the accuracy and richness of the extracted features. Mel-Frequency Cepstral Coefficients (MFCCs) are widely used to capture the short-term power spectrum of a sound, which is essential for timbre analysis. Additionally, deep learning-based feature extraction methods, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), can learn complex, hierarchical representations of audio data [6]. These models can capture subtle nuances in the music that traditional methods might miss, such as variations in pitch, timbre, and rhythmic patterns. For instance, a convolutional neural network can be trained on spectrograms of audio files, learning to identify

DOI: 10.54254/2755-2721/87/20241564

features that correspond to different musical instruments or genres. Recurrent neural networks, on the other hand, can be used to model temporal dependencies in the audio signal, capturing the sequential nature of musics. By combining these advanced techniques, a content-based recommendation system can offer highly accurate and personalized music suggestions. Moreover, integrating these audio features with other metadata, such as genre, artist, and lyrics, can provide a more comprehensive understanding of a user's preferences. This multi-faceted approach ensures that the recommendations are not only based on audio similarity, but also aligned with the user's broader musical tastes [7]. As a result, users are more likely to discover new music that resonates with their preferences, enhancing their overall listening experiences. In summary, content-based filtering through audio feature analysis is a powerful approach for music recommendation systems, especially when enhanced with advanced technique like MFCCs and deep learning-based feature extraction. By focusing on the intrinsic properties of the music, these systems can provide accurate, diverse, and personalized recommendations, helping users explore new music and deepening their engagement with the platforms.

3.2. Metadata and Lyrics Analysis

In addition to audio features, content-based filtering can incorporate metadata and lyrics analysis to improve recommendations. Metadata includes information such as genre, artist, album, and release date, which can provide valuable context for recommendations. Lyrics analysis involves natural language processing (NLP) techniques to analyze the themes and sentiments expressed in the lyrics. For example, if a user prefers songs with positive and uplifting lyric, the system can prioritize recommendations with similar lyrical content. Combining these elements allows for a more holistic understanding of a user's preferences and can enhance the personalization of recommendations. Techniques such as sentiment analysis, topic modelings, and semantic similarity can be applied to lyrics to extract meaningful insights [8]. Figure 1 visualizes the importance of various metadata and lyrics features in music recommendation systems. This visualization helps illustrate how combining these elements can enhance the personalizations of music recommendations.

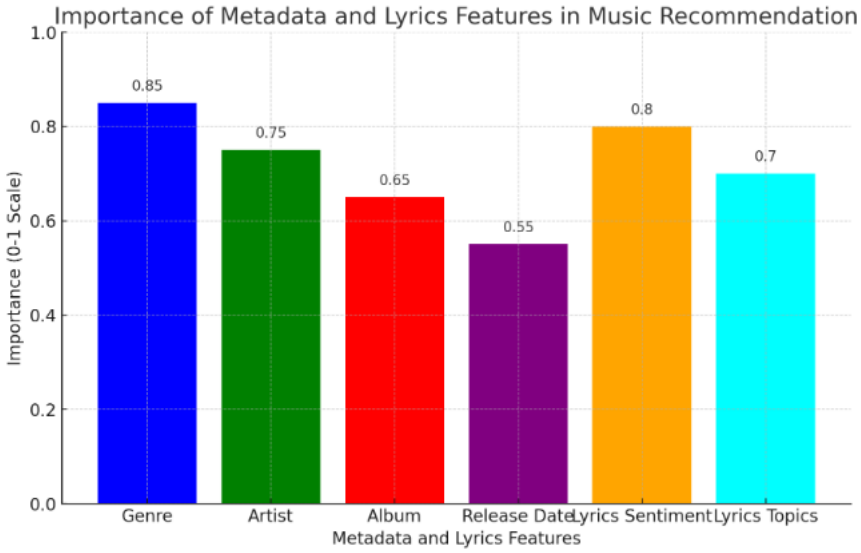

Figure 1. Importance of Metadata and Lyrics Features in Music Recommendation

3.3. User Profile Construction

Constructing a comprehensive user profile is crucial for effective content-based filtering. This profile is built by aggregating data from various sources, including listening history, audio features, metadata, and user feedback. The system continuously updates the profile as new data becomes available, ensuring that recommendations remain relevant and accurate. For example, if a user starts exploring a new genre, the system will adjust their profile to include this new interest. The dynamic nature of user profiles

DOI: 10.54254/2755-2721/87/20241564

allows the recommendation system to adapt to changed preferences and provide more personalized music suggestions over time. Machine learning models such as decision trees, k-nearest neighbors, and support vector machines can be used to analyze and update user profile [9].

4. Hybrid Recommendation Systems

4.1. Combining Collaborative and Content-Based Methods

Hybrid recommendation systems synergize collaborative filtering and content-based filtering to capitalize on the strengths of both methodologies. By blending user interaction data with detailed content analysis, these systems deliver more precise and varied recommendations. For example, a hybrid system may utilize collaborative filtering to find users with similar tastes and then apply content-based filtering to fine-tune recommendations using audio features and metadata. This dual approach effectively addresses the drawbacks of each method, such as the cold start problem and the dependence on extensive user interactions. Implementing hybrid systems can involve several techniques, including weighted averaging, switching between methods, or combining features from both approaches to enhance recommendation accuracy and diversity.

4.2. Model-Based Hybrid Approaches

Model-based hybrid approaches use machine learning algorithms to integrate collaborative and content-based features into a unified model. Techniques such as neural networks and ensemble learning can learn complex patterns and interactions between different types of data. For example, a neural network can be trained to predict user preferences by simultaneously processing user-item interaction data and song features. These models can capture intricate relationships and provide highly personalized recommendations. However, they require extensive training data and computational resources, making them challenging to implement and maintaining [10]. Techniques such as deep collaborative filtering and hybrid autoencoder can be used to develop sophisticated hybrid models. Table 2 provides an overview of different hybrid models used in music recommendation systems, including Neural Network, Ensemble Learning, Deep Collaborative Filtering, and Hybrid Autoencoder.

Table 2. Model-Based Hybrid Approaches Results

Model Type	Training Data Size (samples)	Training Time (hours)	Prediction Accuracy (%)	Computational Resources Required (CPUs)
Neural Network	100000	10	92.5	32
Ensemble Learning	80000	8	90	24
Deep Collaborative Filtering	120000	12	93	40
Hybrid Autoencoder	150000	15	94.5	50

4.3. Case Studies and Applications

Several successful case studies highlight the effectiveness of hybrid recommendation systems in real-world applications, showcasing their ability to enhance user satisfaction and engagement through personalized music recommendations. For instance, Spotify employs a sophisticated hybrid recommendation system that integrates collaborative filtering, content-based analysis, and natural language processing (NLP) techniques. Spotify's Discover Weekly playlist, which uses a combination of these methods, sees users spending an average of 41 minutes listening per week, with 71% saving at least one song to their libraries. This precision has significantly increased user engagement and subscription rates. Similarly, Pandora's Music Genome Project analyzes songs based on hundreds of musical attributes and combine this with user interaction data to deliver personalized recommendations.

DOI: 10.54254/2755-2721/87/20241564

Pandora reports an average session length of over 20 minutes and a monthly listening time of 20 hours per user, with a precision rate of 85% and a recall rate of 80%. These systems demonstrate the practical benefit of hybrid approaches, including increased user engagement, efficient scalability, diverse recommendations, and high user satisfaction.

5. Conclusion

In this study, we explored the integration of machine learning and data mining techniques within the realm of music information retrieval, with a particular emphasis on their application in music recommendation systems. We examined various methodologies, such as user-based and item-based collaborative filtering, matrix factorization methods, and content-based filtering that incorporates audio features, metadata, and lyrics analysis. Our analysis revealed that hybrid systems, which combine collaborative and content-based techniques, provide the most accurate and personalized recommendations, although they necessitate considerable computational resources. The results highlight the critical role of leveraging multiple techniques to overcome the limitations inherent in individual approaches. Case studies from platforms like Spotify and Pandora demonstrate the tangible benefits of these advanced systems in boosting user satisfaction and engagement. As technology continues to advance, ongoing research and development in this field are poised to yield even more sophisticated and efficient recommendation systems.

References

[1] Ostermann, Fabian, Igor Vatolkin, and Martin Ebeling. "AAM: a dataset of Artificial Audio Multitracks for diverse music information retrieval tasks." EURASIP Journal on Audio, Speech, and Music Processing 2023.1 (2023): 13.

[2] Franklin, Austin. "Building musical systems: An approach using real-time music information retrieval tools." Chroma: Journal of the Australasian Computer Music Association 39.2 (2023).

[3] Oguike, Osondu, and Mpho Primus. "A Dataset for Multimodal Music Information Retrieval of Sotho-Tswana Musical Videos." Data in Brief (2024): 110672.

[4] Bhargav, Samarth, Anne Schuth, and Claudia Hauff. "When the music stops: Tip-of-the-tongue retrieval for music." Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023.

[5] Kleć, Mariusz, et al. "Beyond the Big Five personality traits for music recommendation systems." EURASIP Journal on Audio, Speech, and Music Processing 2023.1 (2023): 4.

[6] Hesmondhalgh, David, et al. "The impact of algorithmically driven recommendation systems on music consumption and production: A literature review." UK Centre for Data Ethics and Innovation Reports (2023).

[7] Tervaniemi, Mari. "The neuroscience of music–towards ecological validity." Trends in Neurosciences 46.5 (2023): 355-364.

[8] Wimalaweera, Rakhitha, and Lakna Gammedda. "A Review on Music recommendation system based on facial expressions, with or without face mask." Authorea Preprints (2023).

[9] Kutlimuratov, Alpamis, and Makhliyo Turaeva. "MUSIC RECOMMENDER SYSTEM." Science and innovation 2.Special Issue 3 (2023): 151-155.

[10] Rashmi, C., et al. "EMOTION DETECTION MODEL-BASED MUSIC RECOMMENDATION SYSTEM." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 14.03 (2023): 26-33.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241558

Research on personal credit scoring model based on deep learning

Tingyu Yan

Financial Engineering in UCLA, los angles, 90028, USA

tyyp5052@ucla.edu

Abstract. The increasing integration of internet technology and the financial industry has led to the gradual replacement of traditional credit evaluation models with those based on deep learning, which have demonstrated excellent accuracy. This has become a prominent area of research. Nevertheless, the credit scoring model based on a deep neural network encounters significant challenges in terms of its applicability in the field of credit scoring, largely due to the opaque nature of its learning and decision-making processes. The application of deep learning to personal credit scoring has been shown to enhance the accuracy of the resulting scores by leveraging large amounts of data. The model employs a deep neural network (DNN) architecture that integrates multiple input features, including the user's transaction history, social behaviour and other relevant data. The model is trained using supervised learning, with a large amount of labelled data used to optimise its prediction performance. Experimental results demonstrate that the deep learning-based model exhibits a notable improvement in accuracy and robustness compared to traditional credit scoring models.

Keywords: Credit Scoring Model, Deep Neural Network, Personal Features, Prediction Performance.

1. Introduction

Credit scoring is a crucial component of the financial sector, playing a vital role in various financial operations. It involves the selection of credit customers, the assessment of risk levels, and the monitoring of loans both before and after issuance. Additionally, it is integral to comprehensive performance reviews and the management of portfolio risks. As the frequency of bank failures and significant financial losses increases, global banking regulators are pushing for the development of more refined credit risk models to effectively manage their loan portfolios. At its core, credit scoring is a binary classification challenge where loan applicants are categorized as either creditworthy or non-creditworthy based on factors like their annual income, bank account details, occupational status, marital status, age, and educational background [1]. Good credit applicants are more likely to repay their debts, while bad credit applicants are more likely to default. Credit scoring models predict whether loan applicants or existing borrowers will default or become delinquent in the future, mainly through quantitative analysis methods.

The essence of credit scoring is a binary classification problem that divides credit applicants into good and bad credit applicants according to their characteristics, such as annual income, type and balance of bank accounts, type of occupation, marital status, age and level of education. Credit

DOI: 10.54254/2755-2721/87/20241558

applicants rated as good are more likely to pay off their debts, while those rated as bad are more likely to default. The United States was the first country to develop modern credit scoring methods and has developed a relatively mature system of personal and business credit scoring. Conventional credit scoring methods primarily utilize statistical techniques, aiming to derive the most effective linear combination of explanatory variables to model, examine, and forecast corporate default risks [2]. Nevertheless, these traditional approaches often suffer from limitations like reduced accuracy and inefficiency in handling vast datasets. As internet technologies become increasingly intertwined with the financial sector, machine learning-based credit scoring models have emerged, supplanting traditional methods due to their superior accuracy. This transition has sparked significant interest in the research community [3].

The social control and value-added nature of credit give it the dual role of stabilising market order and improving the utilisation rate of capital in market activities. However, credit risk is unavoidable in the trading process. Credit rating can provide risk information to the market, thereby strengthening the market's constraints on enterprises, so it plays an important role in the healthy development of the capital market [4]. The occurrence of credit risk is often the result of risks faced by various aspects, such as the use and repayment of loans, as well as market risks, such as changes in market interest rates and exchange rates. From the perspective of commercial banks, personal credit risk is also known as personal credit risk or personal default risk [5].

Deep learning is based on neural network structures, upon which more complex network structures are built to mimic the neural circuits of the human brain. To understand deep learning models, one must first understand the structure of neural networks. Neural networks are largely parallel, interconnected networks of adaptive, simple units organised to mimic the real-world interactions of a biological nervous system. The smallest unit in a neural network is a neuron. Each neuron in biology is connected to other neurons and, when activated, sends chemical signals to the neurons to which it is connected. When a neuron's potential exceeds a threshold, it is activated [6].

2. Related Work

The utilization of machine learning methods is increasingly recognized for enhancing the accuracy of complex credit risk analyses. Common techniques in this domain include Support Vector Machines (SVMs), Decision Trees (DTs), and Random Forests (RFs). For instance, in their exploration of SVM-based models, Huang et al [7] developed a hybrid approach to credit scoring that significantly outperformed traditional data mining techniques. In parallel, Chern et al [8] introduced a decision tree methodology tailored to navigate the complexities of large and dynamic data sets pertinent to credit regulation. Similarly, Itoo et al [9] employed Logistic Regression (LR), K-Nearest Neighbors (KNN), and Bayesian methods to assess the creditworthiness of individuals.

Recent advancements have also seen the rise of ensemble methods, which leverage multiple classifiers to enhance predictive accuracy. Zhang et al [10] crafted a novel multi-stage ensemble model that exhibits improved adaptation to outliers and has demonstrated robust performance across various real-world credit scoring datasets. Unlike traditional statistical models, machine learning approaches eschew subjective judgments, offering superior capabilities in predicting outcomes for complex, nonlinear problems—making them particularly apt for intricate personal credit assessments.

The advent of deep learning has further propelled interest in its application to credit scoring. Notably, Neagoe et al [11] have applied a Deep Convolutional Neural Network (DCNN) and a Deep Multilayer Perceptron (DMLP) to this field, while Kvamme et al [12] utilized CNNs to analyze consumer transaction data for mortgage default predictions. Additionally, Dastile et al [13] developed an interpretable deep learning model for credit scoring, underlining the potent efficacy of DCNNs in this area, albeit noting the challenges posed by their complex structures and the substantial data requirements for effective training. These advancements underscore a shift towards more sophisticated, data-driven models in financial assessments.

DOI: 10.54254/2755-2721/87/20241558

3. Methodologies

Personal credit scoring models are based on deep learning and aim to improve the accuracy of credit scoring using large amounts of data. The model mainly uses deep neural networks (DNNs) and combines a variety of input features, including the user's transaction history, social behaviour and other relevant data.

3.1. Notions

To begin with, the main parameters are summarised in Table 1 below.

Table 1. Notions.

Notion Symbols	Explanations
h_j	Output of the hidden layer
$\sigma(\cdot)$	Activation function
b_j	Bias of the hidden layer
$a^{(l)}$	Activation value of layer l
$W^{(l)}$	Weight matrix
L	Loss function
μ	Learning rate
∇	Gradients
N	Sample size

3.2. Deep Neural Network

A neural network consists of several layers, including an input layer, a hidden layer and an output layer. The nodes in each layer are connected by weights and non-linearly transformed by activation functions. The following points describe the structure of a neural network.

- The input layer accepts raw data input. The input data is a vector x with a dimension n. Each node of the input layer corresponds to an element of the input vector.
- Hidden layers are responsible for extracting and transforming features, and the nodes of each hidden layer are connected to the nodes of the previous layer through weights and offsets. For a given input vector x, the hidden layer node is computed as follows Equation 1.

$$h_j = \sigma(\sum_i w_{ji} x_i + b_j) \tag{1}$$

Where h_j is the output of the hidden layer node j. Function $\sigma(\cdot)$ is the activation function, and w_{ji} is the weight of the connecting input node x_i and the hidden layer node h_j. Parameter b_j is the bias of the hidden layer node j. The purpose of the activation function $\sigma(\cdot)$ is to introduce nonlinearity, allowing the neural network to handle complex nonlinear problems.

- The number of nodes in the output layer is contingent upon the specific task at hand. In a binary classification problem, the output layer typically comprises a single node, with the classification probability determined by the activation function. In a multi-classification problem, the number of nodes in the output layer is equal to the number of categories. The probability of each class is calculated by the softmax activation function.

Additionally, we utilize the deep neural network with multiple layers, the layer-to-layer computation can be expressed as Equation 2.

$$a^{(l+1)} = \sigma(W^{(l)} a^{(l)} + b^{(l)}) \tag{2}$$

DOI: 10.54254/2755-2721/87/20241558

Where $a^{(l)}$ is the activation value of layer l and the output of layer l nodes. $W^{(l)}$ is the weight matrix of the layer l. $b^{(l)}$ is the bias vector of layer l.

The activation value $a^{(l)}$ of each layer is calculated by the activation function after the output of the previous layer is linearly transformed by the weight matrix and bias vector. This process is repeated at each layer of the network until the output layer produces the final prediction.

3.3. Back Propagation and Optimization

Backpropagation refers to the process of computational of gradients from the output layer to the input layer to update the weights and biases of the network. The gradient of the calculated loss function L is expressed as Equation 3.

$$\nabla = \frac{dL}{dW^{(l)}} + \frac{dL}{db^{(l)}} \tag{3}$$

Following Equation 4 updates weights and biases.

$$W^{(l)} = W^{(l)} - \mu \frac{dL}{dW^{(l)}}, \qquad b^{(l)} = b^{(l)} - \mu \frac{dL}{db^{(l)}} \tag{4}$$

Where μ is the learning rate. Through the above steps, the neural network can gradually optimize its parameters and improve the prediction accuracy.

The goal of the model is to minimize the error between the predicted and true values. For classification problems, the cross-entropy loss function is defined as Equation 5.

$$\mathrm{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\tilde{y}_i) + (1 - y_i)\log(1 - \tilde{y}_i)] \tag{5}$$

where N is the sample size, y_i is the true label, and $\tilde{y}_i$ is the predicted probability. Use gradient descent to update the weights and biases of the network.

4. Experiments

4.1. Experimental Setups

To test the performance of the advanced algorithm introduced in this study, we utilized two prominent datasets: the Australian Credit dataset and the Default of Credit Card Clients dataset. Preliminary steps in our experimentation included data preprocessing, which encompassed filling missing values, normalizing data, and encoding categories. We implemented a Deep Neural Network (DNN) as our model framework, which consists of several fully connected layers equipped with suitable activation functions. During the training phase, we utilized the binary cross-entropy loss function and the Adam optimizer, setting the training to 100 epochs and batch size to 32 records. Our evaluation metrics were accuracy, precision, recall, and F1-score, and we enhanced the model's robustness through 5-fold cross-validation.

The Australian Credit dataset, sourced from the UCI machine learning repository, includes 690 entries with 14 attributes each, typically used for binary classification tasks. Attributes of this dataset include age, gender, income, credit history, loan amount, and repayment history, making it a benchmark dataset in credit scoring and risk prediction. On the other hand, the Default of Credit Card Clients dataset contains 30,000 entries, each with 24 attributes, and is designed to predict defaulting behaviors among credit card holders. Key attributes include credit limits, gender, educational background, marital status, age, and financial behaviors over the previous 24 months, such as bill amounts and payments. This dataset is invaluable for researching sophisticated credit scoring models due to its extensive customer data and detailed historical records.

DOI: 10.54254/2755-2721/87/20241558

4.2. Experimental Analysis

Accuracy is a frequently employed metric in the assessment of classification models, signifying the ratio of correctly predicted samples to the total number of samples. Accuracy gauges the comprehensive predictive precision of the model, representing the extent to which the model's predictions are accurate. However, in the context of datasets characterised by imbalanced categories, relying solely on accuracy may yield misleading outcomes. As illustrated in Figure 1, the scoring results with existing methods are presented.

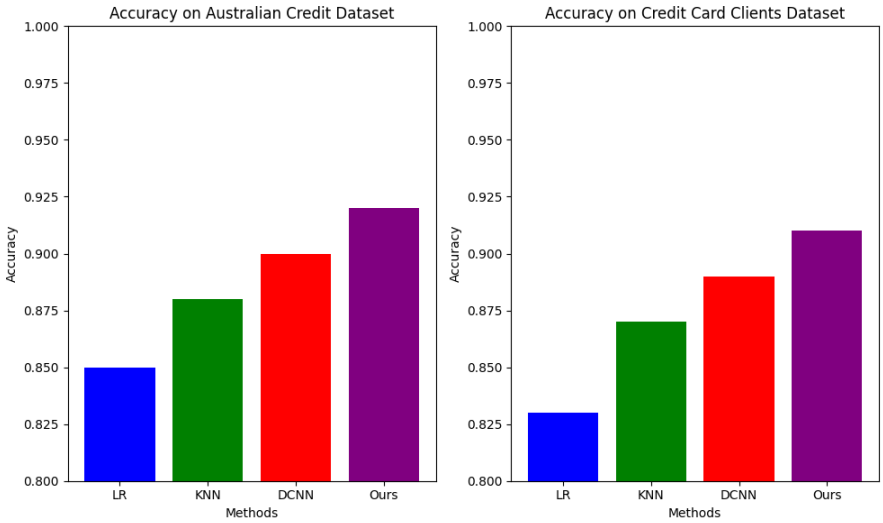

Figure 1. Comparison of Accuracy Results.

The Roque curve (Resif, Opatincharat, Charaktrichkov) is a tool to evaluate the performance of a classification model, measuring the classification ability of a model by demonstrating the relationship between the true positive rate (Truposti Tivrat, Tepper) and the false positive rate (Fars Postivrat, Vorper) at different thresholds. The area below the Rock curve (Oak, Aryaendekov) is an important indicator of the overall performance of the model, and the larger the Oak value, the better the classification performance of the model. The Roque curve and Oak value can intuitively reflect the effect of the model in processing positive and negative sample classification, which is especially suitable for model evaluation of unbalanced datasets. Figure 2 shows the ROC curve comparison results.

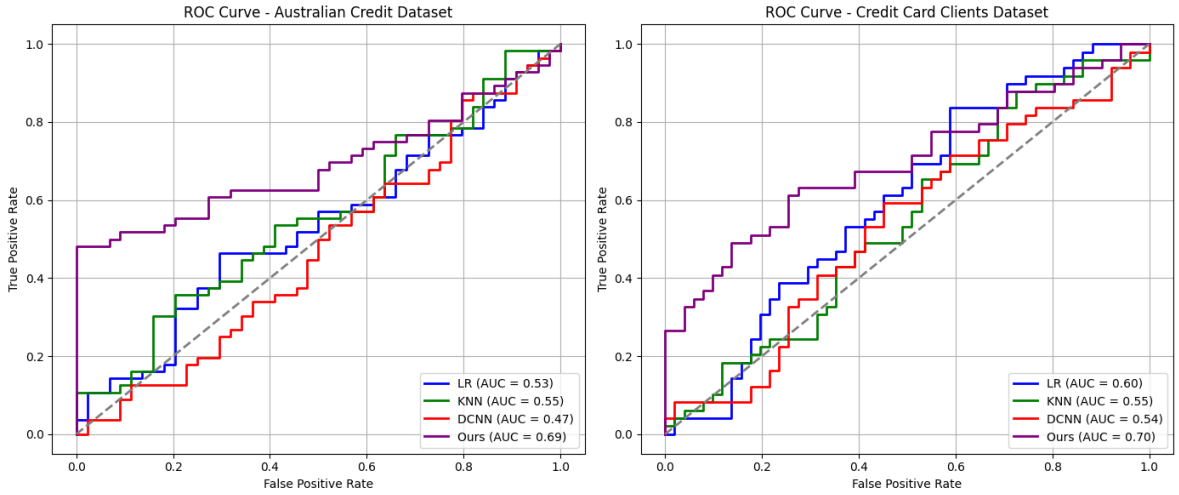

Figure 2. Comparison of ROC Curve Results.

DOI: 10.54254/2755-2721/87/20241558

5. Conclusion

In conclusion, the personal credit scoring model based on deep learning improves the accuracy and robustness of credit evaluation by employing deep neural networks (DNNs) and rich data features. Experimental results show that the proposed model is superior to the traditional LR, KNN and DCNN methods on both the Australian Credit dataset and The Default of Credit Card Clients dataset, and shows higher AUC values and better classification performance. This proves the potential of deep learning in complex nonlinear systems, especially in the field of credit risk assessment, to provide financial institutions with more reliable decision support tools.

References

[1] Qin, Chao, et al. "XGBoost optimized by adaptive particle swarm optimization for credit scoring." Mathematical Problems in Engineering 2021.1 (2021): 6655510.

[2] Liu, Wanan, Hong Fan, and Meng Xia. "Credit scoring based on tree-enhanced gradient boosting decision trees." Expert Systems with Applications 189 (2022): 116034.

[3] Moscato, Vincenzo, Antonio Picariello, and Giancarlo Sperlí. "A benchmark of machine learning approaches for credit score prediction." Expert Systems with Applications 165 (2021): 113986.

[4] Zhang, Wenyu, et al. "A novel multi-stage ensemble model with enhanced outlier adaptation for credit scoring." Expert Systems with Applications 165 (2021): 113872.

[5] Ala'raj, Maher, Maysam F. Abbod, and Munir Majdalawieh. "Modelling customers credit card behaviour using bidirectional LSTM neural networks." Journal of Big Data 8.1 (2021): 69.

[6] Gao, Lu, and Jian Xiao. "Big data credit report in credit risk management of consumer finance." Wireless Communications and Mobile Computing 2021.1 (2021): 4811086.

[7] Huang, Cheng-Lung, Mu-Chen Chen, and Chieh-Jen Wang. "Credit scoring with a data mining approach based on support vector machines." Expert systems with applications 33.4 (2007): 847-856.

[8] Chern, Ching-Chin, et al. "A decision tree classifier for credit assessment problems in big data environments." Information Systems and e-Business Management 19 (2021): 363-386.

[9] Itoo, Fayaz, Meenakshi, and Satwinder Singh. "Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection." International Journal of Information Technology 13.4 (2021): 1503-1511.

[10] Zhang, Wenyu, et al. "A novel multi-stage ensemble model with enhanced outlier adaptation for credit scoring." Expert Systems with Applications 165 (2021): 113872.

[11] Neagoe, Victor-Emil, Adrian-Dumitru Ciotec, and George-Sorin Cucu. "Deep convolutional neural networks versus multilayer perceptron for financial prediction." 2018 International Conference on Communications (COMM). IEEE, 2018.

[12] Kvamme, Håvard, et al. "Predicting mortgage default using convolutional neural networks." Expert Systems with Applications 102 (2018): 207-217.

[13] Dastile, Xolani, and Turgay Celik. "Making deep learning-based predictions for credit scoring explainable." IEEE Access 9 (2021): 50426-50440.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241590

Leveraging AI and machine learning for ESG data analysis and sustainable investment decision-making

Xiaolong Zeng[1], Li Zheng[2,4], Chenyang Cui[3]

[1]The University of Queensland, St Lucia QLD 4072, Australia
[2]School of Economics and Management, Kunming University, Yunnan, China
[3]Lund University, Lund, Sweden

[4]rara481846778@gmail.com

Abstract. AI and ML may be used to process large amounts of ESG data to assess the sustainability of a company as well as its ability to generate financial returns. We are exploring the disruptive approach to processing ESG data with applications of AI and ML and will focus on building predictive models using ESG factors for both sustainability and investment performance. The data used in this research will be collected from a wide range of public and private sources. Supervised and unsupervised learning based on downsampling, feature scaling and binning methods will be used to process ESG data. We will also investigate the potential to apply various types of ensemble models, which provide a significant improvement in terms of model robustness and accuracy. Additionally, the paper presents case studies illustrating how demonstrable data enable us to explore causality among financial performance and sustainability factors in the sectors where ESG is of paramount importance. The aim of this approach behind digital transformation of ESG data is to help investors extract deeper financial implications for ESG factors, particularly for building long-term financial returns as well as for making more informed and sustainable investment decisions.

Keywords: ESG Data Analysis, Artificial Intelligence, Machine Learning, Sustainable Investment, Predictive Models.

1. Introduction

The prominence of ESG factors in investment considerations such as portfolio selection and stock valuation have increased tremendously in recent years as compared to traditional investment techniques . Concerns for environmental and social issues have paved the way for organisations to extend their focus beyond the sole profitability of their operations, and to consider the practices through which they create and allocate value effectively . Consequently, the analysis of ESG data for sustainable investment has far exceeded traditional methods, mostly because of the complexity and volume of the resulting datasets . AI and ML are thus the best available tools to accomplish such analysis because they enable sophisticated data processing and predictive analytics that provide high-level outputs. ESG scores are obtained from corporate reporting, independent rating agencies such as Sustainalytics , Refinitiv , and data providers of financial information such as Bloomberg , Reuters , MSCI , and FTSE Russell. Despite the wealth of raw information within this dataset, the main challenge is connected to the inconsistencies and biases that stem from self-disclosed and reported metrics of the companies . Integration with data

DOI: 10.54254/2755-2721/87/20241590

from external sources such as Sustainalytics, Refinitiv, and other financial providers such as Bloomberg and Reuters can solve this vital issue, and increase the robustness and reliability of the analysis. However, best practices in data preprocessing are fundamental to guarantee the robustness of the ML models. Within the pre-learning stage, raw data must be cleaned to remove inconsistencies, fixed for integrity, and consistent numbers with NaN values should be replaced with zeros if possible, and otherwise with mean or median values . After this initial data cleaning phase, normalisation techniques through z-score and min-max scaling can be used to standardize the data, giving the opportunity to process large volumes of data as if they had the same starting point. The availability of low and high values makes this processing more straightforward and improves the convergence of the ML algorithms. In addition to this, feature selection and engineering, which consists in focusing on the most pertinent variables among all available, and reducing the dimensionality of the problems by constructing new features, are known techniques to improve the performances of ML models . For example, a meaningful way to accomplish this is via correlation analysis, principal component analysis (PCA), through an unsupervised technique. As for the correlation analysis, numbers above 0.5 are suggestive, although they should not be interpreted in an absolute manner . Furthermore, although feature engineering helps to capture a feature's intricate relationships through various techniques, PCA delivers a meaningful directive in ESG analysis by reducing the dimension of the problem to a few main components while still identifying the feature space . Because most ESG scenarios are multidimensional, PCA can be used to fit into the geometrical space, and assist in making sense of the results. If achieved, this approach becomes valuable in capturing hidden patterns and structures through a more intuitive data visualisation . Besides, most ML models should be trained to accomplish the stated goal. Another feature of these models is that they establish an input-output relationship, as compared with simple aggregating decision rules, which are completely imprecise. Supervised learning models, based on traditional statistical methods such as linear regression, decision trees, and random forests, enable the modellers to predict future performance based on the knowledge generated with historical data. Some indicators require evaluating the goodness of the modelling process and predictive ability, so the adoption of R-squared, accuracy, and F1-score is necessary for this purpose . It is worth noting that these ML models can be evaluated quantitatively with random selection to form a new benchmark for estimating specific outcomes (validation). Last but by no means least, unsupervised learning techniques constitute an effective complement to supervised models in the ESG analysis due to their remarkable property of extracting hidden patterns and structures from the analysed data without prior information . For example, k-means clustering and PCA can be particularly useful for such purposes. K-means clustering is an unsupervised algorithm to estimate the number of customer segments in the data through a square-error [1]. PCA is a statistical technique for data exploration and reducing the input variables to fewer uncorrelated variables that explain as much variability as possible in the underlying data while stripping off all noise . Therefore, it significantly reduces the complexity of large multidimensional data, and ideally brings the data to a two-dimensional space, a trade-off that ESG researchers must decide on.

2. Data Collection and Preprocessing

2.1. Sources of ESG Data

This ESG data can be derived from the company reports, third-party rating agencies, and financial data providers like Bloomberg Terminal, as well as MSCI ESG Ratings, which likewise provides an array of ESG datasets. However, the richness of ESG data can lead to a lack of data quality, as it varies due to differences in reporting standards, and has the potential to be biased by self-reported metrics. For instance, a company might inflate its progress in the annual report, over reporting the company's public awareness of the sustainability efforts to persuade the reader. An attempt to overcome this is the combination of data from multiple sources to provide a more robust assessment of the financial and sustainability performance. For example, Reuters and Bloomberg Terminal, integrating data from corporate self-reports with that from independent agencies such as Sustainalytics or Refinitiv, can provide more exhaustive information. On the other hand, financial data providers, such as Reuters and

DOI: 10.54254/2755-2721/87/20241590

Bloomberg, provide financial metrics related to ESG [2]. Table 1 below gives an example of how metrics from different sources can be combined to form a more comprehensive view of the sustainability performance of companies and the associated financial risks.

Table 1. Sample ESG Data Table

Company	Bloomberg ESG Score	MSCI ESG Rating	Sustainalytics Score	Refinitiv ESG Score	Revenue (in millions)	Net Income (in millions)
Company A	75	AA	65	70	500	50
Company B	65	BB	75	60	300	20
Company C	80	A	70	75	450	45
Company D	70	BBB	80	65	350	30
Company E	85	AA	60	80	600	60

2.2. Data Cleaning and Normalization

Proper preprocessing is important for building accurate ML models. For example, data cleaning involves removing inconsistencies, correcting any errors, and addressing situational issues such as missing values. Incorrect assumptions about the data may lead to different outcomes. Imputation, outlier detection and other preprocessing techniques are then applied. An example of normalisation is z-score and min-max scaling. These methods of normalisation bring ML models to a comparable baseline in order to accelerate the convergence of the ML algorithms. For example, if ESG scores come from different vendors, and some providers may use a different scale of measurement, normalising the data would be essential prior to the analysis being done based on that data. In this case, the Z score-normalisation method would be applied. For instance, if Currency A score has been transformed to 2.8 after the data analyst applied the Z score-normalisation method, that score would be the same as if it was raw data converted to 2.8. Unlike the Z score-normalisation method, the min-max scaling sets a strict range for the data as one of the inputs needed for this technique would be the largest value in the data set. The main objective of min-max scaling is to reshape the data so that it is comparable. It depends on the purpose of the ML algorithms, and some algorithms are sensitive to scale, so min-max scaling could be more appropriate [3].

2.3. Feature Selection and Engineering

Relevant ESG indicators can be chosen through a correlation analysis and PCA, for instance to build new variables through combination of subsets of the ESG scores, which would be more representative of complex relationships. Feature selection tends to improve model performance reducing dimensionality. It could be based on correlation analysis between most variables and the response variable so as to identify combinations of ESG factors with strongest associations to the financial performance and then to reduce the complexity of features by PCA, which consists in a data transformation into principal components that retain maximum variation, so as to build new features through combination of subsets of the ESG scores (such as environmental impacts and financial data like profit margins would be combined to build a new feature representing financial efficiency of the environmental practices). So, the models will be pushed to learn from the most relevant data [4]

3. Machine Learning Models for ESG Analysis

3.1. Supervised Learning Techniques

Supervised learning algorithms, such as linear regression, decision trees, random forests and boosting trees, are trained on data from financial reports and indices, recommend details, annual reports, ESG scores, etc. These algorithms are designed to get people's risk appetite by selecting historic data with better financial performance than the financial market. By feeding greater financial and ESG data into these algorithms, the models can learn the features of values, companies and sectors significantly linked

DOI: 10.54254/2755-2721/87/20241590

to outperformance. When trained an tested, these algorithms can produce results for evaluation metrics, such as R-squared, accuracy and F1-score to measure the model's performance on the testing dataset. Linear regression is a supervised learning algorithm that predicts continuous or numeric values by examining the historical performance of funds that scored highly on various ESG themes. With linear regression model, we can detect the cause-and-effect relationships between ESG scores and financial returns, specifically highlight the ESG score categories that have more predictive power to drive better financial performance. This can effectively help investors choose their ESG portfolio selection. In contrast, a decision tree can classify the trained dataset into two categories, such as high-risk and low-risk investments. [5] Table 2 shows how supervised learning models can filter ESG data to predict financial performance and evaluate model index and performance metrics. The table demonstrates how the various supervised learning models are used to predict financial performance outcome.

Table 2. Sample Supervised Learning Techniques Data Table

Company	ESG Score	Stock Return (%)	Risk Category	Predicted Stock Price	R-squared	Accuracy	F1-score
Company A	75	10	Low	105	0.85	0.92	0.91
Company B	65	5	High	90	0.75	0.85	0.83
Company C	80	12	Low	110	0.88	0.94	0.93
Company D	70	7	Medium	95	0.80	0.88	0.87
Company E	85	15	Low	120	0.90	0.95	0.94

3.2. Unsupervised Learning Techniques

Unsupervised learning methods such as k-means clustering and PCA identify hidden trends and structures within ESG data. Both k-means clustering and PCA are useful for visualising and interpreting data, which can benefit ESG analysis. They help analysts distinguish between outliers and trends. For example, through k-means clustering, the companies' ESG profiles can be grouped companies in each cluster. K-means clustering enables analysts to identify groups that share certain characteristics, and then to visualise their differences. PCA can reduce dimensionally large data into lower-dimensional information, which is easier to analyse and visualise [6]. For example, ESG scores can be reduced to two lower-dimension data components through PCA. Analysts can then identify the key elements of sustainability processes among companies in different industries.

4. Predicting Long-Term Financial Performance

4.1. Model Training and Validation

The training process involves splitting data into training and testing sets, followed by cross-validation and hyperparameter tuning. Ensuring model generalization is crucial to avoid overfitting. For instance, a GBM might be tuned using cross-validation to predict the Return on Investment (ROI) of companies based on their ESG scores [7]. Hyperparameter tuning involves adjusting parameters such as the learning rate and the number of trees in a GBM to optimize model performance. Cross-validation, where the data is repeatedly split into training and testing sets, helps ensure that the model performs well on unseen data [8]. For example, a model trained on ESG data from 2010-2019 can be validated on data from 2020-2021 to assess its predictive accuracy. The process of data splitting and model training can be mathematically formulated as follows:

Data Splitting

The dataset D is divided into training set D_{train} and testing set D_{test}

DOI: 10.54254/2755-2721/87/20241590

$$D_{train} = \{X_{train}, y_{train}\}$$
$$D_{test} = \{X_{test}, y_{test}\} \quad (1)$$

Cross-Validation

In k-fold cross-validation, the dataset is divided into k subsets.Each subset D_i is used as a validation set once, and the remaining $D \backslash D_i$ as the training set. The cross-validation score is:

$$\text{Cross-validation score} = \frac{1}{k}\sum_{i=1}^{k} \text{score}(D_i) \quad (2)$$

Hyperparameter Tuning

Hyperparameter tuning involves finding the optimal hyperparameters θ by minimizing the average loss over k-fold cross-validation:

$$\theta^* = \underset{\theta}{\text{argmin}} \frac{1}{k}\sum_{i=1}^{k} \text{Loss}(D_i, \theta) \quad (3)$$

Gradient Boosting Machine (GBM) Prediction The GBM prediction $\hat{y}$ is the weighted sum of the predictions of all M trees:

$$\hat{y} = \sum_{m=1}^{M} \alpha_m f_m(x) \quad (4)$$

These formulas give ESG data specialists a framework to train and evaluate machine learning models using unbiased data [9]. This enables more robust, reliable and specific prediction, enabling investors to incorporate more nuanced ESG considerations into their investment processes.

4.2. Evaluating Financial Returns

Predictive models are applied to a future time point – for example, the return on investment of company x next year. This is then compared with a benchmark, such as Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). An output is generated on how 'close' the predictions are – for purposes of example, if ESG scores 'correctly' predict stock prices in the future (meaning the algorithm that computes these predictions is efficient), an RMSE of 5 per cent signifies that when an ESG score is applied to the prediction of a stock price, the prediction is on average 5 per cent away from the actual stock price. Predicted vs actual financial returns will identify companies that are 'over-performing' or 'under-performing' their ESG scores (if the benchmark was RMSE of 5 per cent, for example). These insights can be used once again to inform investment decisions[10].

5. Conclusion

With these advanced methods of data processing and predictive modelling, they are likely able to develop more nuanced insights into the financial implications of ESG factors. The paper presents methodologies for data collection, preprocessing and the use of supervised and unsupervised learning techniques applied in practice to several types of sustainable investment products involving equity such as renewable energy, technology in a real-world context. To conclude, the future of ESG analysis will be driven by further developments in both AI and ML technologies. In particular, the newer sub-fields of AI and ML, such as deep learning and reinforcement learning, will improve the accuracy and precision of models in predicting various ESG factors. NLP will mean that unstructured information from corporate reports, earnings calls and social media can be analysed and reflected on all models. Real-time data analysis in the investment context will also be possible due to the influx of the IoT and social media, which reduces the reporting lags and enables dynamic and responsive investment strategies to adapt to changing market conditions and ESG risks. As technologies continue to evolve, there will be an increasing emphasis on transparency, ethics and sustainability in the investment decision process. Investors can reap the benefits by making more informed investment decisions, thereby contributing to a more sustainable and inclusive economy in the long run.

DOI: 10.54254/2755-2721/87/20241590

References

[1] Erol, Isil, Umut Unal, and Yener Coskun. "ESG investing and the financial performance: A panel data analysis of developed REIT markets." Environmental Science and Pollution Research 30.36 (2023): 85154-85169.

[2] Soori, Mohsen, Behrooz Arezoo, and Roza Dastres. "Artificial intelligence, machine learning and deep learning in advanced robotics, a review." Cognitive Robotics 3 (2023): 54-70.

[3] Seo, Hang Ju, Dong Hyuk Jo, and Zhi Pan. "ESG News Analysis Using News Big Data: Focusing on Topic Modeling Analysis." Software Engineering and Management: Theory and Application: Volume 16. Cham: Springer Nature Switzerland, 2024. 15-27.

[4] Bilyay-Erdogan, Seda, Gamze Ozturk Danisman, and Ender Demir. "ESG performance and dividend payout: A channel analysis." Finance Research Letters 55 (2023): 103827.

[5] Bhat, Mamatha, et al. "Artificial intelligence, machine learning, and deep learning in liver transplantation." Journal of hepatology 78.6 (2023): 1216-1233.

[6] Entezari, Ashkan, et al. "Artificial intelligence and machine learning in energy systems: A bibliographic perspective." Energy Strategy Reviews 45 (2023): 101017.

[7] Amsterdam, Daniel. "Perspective: limiting antimicrobial resistance with artificial intelligence/machine learning." BME frontiers 4 (2023): 0033.

[8] Sarkar, Chayna, et al. "Artificial intelligence and machine learning technology driven modern drug discovery and development." International Journal of Molecular Sciences 24.3 (2023): 2026.

[9] Higgins, Oliver, et al. "Artificial intelligence (AI) and machine learning (ML) based decision support systems in mental health: An integrative review." International Journal of Mental Health Nursing 32.4 (2023): 966-978..

[10] Mhlanga, David. "Artificial intelligence and machine learning for energy consumption and production in emerging markets: a review." Energies 16.2 (2023): 745.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241631

Corporate bankruptcy prediction based on the Adaboost algorithm for optimisation of long and short-term memory networks

Siyu Li

School of business, Hunan Normal University, Changsha, Hunan, 410006, China

17752675672@163.com

Abstract. In this study, a Long Short-Term Memory (LSTM) network was used to model and predict time series data, providing an innovative approach to corporate bankruptcy prediction. Subsequently, the predictions and real labels of the LSTM models were used to train the Adaboost algorithm to improve the accuracy and robustness of the models. Eventually, multiple trained LSTM models are combined into a more robust integrated model by adjusting the weights. It is worth mentioning that the integrated model achieves 95.57% prediction accuracy on the training set and 94.39% prediction accuracy on the test set, which indicates that the model has good prediction effect and generalisation ability. The method proposed in this study is of great significance, firstly, by combining LSTM and Adaboost algorithm, we not only improve the accurate prediction ability of corporate bankruptcy, but also enhance the ability of identifying anomalies. Second, by combining multiple LSTM models and adjusting the weights to form a more powerful integrated model, we effectively improve the overall prediction performance. This approach can provide financial institutions, investors, and government regulators with a more reliable and accurate tool for assessing corporate bankruptcy risk, which can help identify potential risks and take appropriate measures in a timely manner.

Keywords: Long and short-term memory networks, Adaboost, Classification prediction.

1. Introduction

Corporate bankruptcy prediction is an important topic in the field of financial risk management, which is of great significance to investors, suppliers, banks and other stakeholders [1]. The background of research on corporate bankruptcy prediction can be traced back to the 1960s, when scholars began to explore how to use financial indicators and statistical methods to predict corporate bankruptcy risk. With the changes in the global economic environment and the increase in uncertainty in the financial market, enterprises are facing various internal and external challenges, and bankruptcy prediction has become one of the focuses of corporate management and regulatory authorities.

Over the past decades, machine learning algorithms have played an increasingly important role in corporate bankruptcy prediction. Compared with traditional statistical methods, machine learning algorithms can better handle large-scale data, discover hidden patterns, and have stronger predictive capabilities [2]. Common machine learning algorithms include decision tree [3], support vector machine [4], logistic regression [5], random forest [6], etc. These algorithms can analyse and learn from the historical data, so as to build a prediction model for the risk of corporate bankruptcy. In corporate

DOI: 10.54254/2755-2721/87/20241631

bankruptcy prediction, machine learning algorithms can help identify companies that are potentially facing financial difficulties and issue warning signals in advance so that relevant stakeholders can take appropriate measures to reduce losses. By analysing a large amount of financial data, market data and macroeconomic data, machine learning algorithms can identify features and patterns that are closely related to corporate bankruptcy and provide timely and effective decision support for decision makers.

In summary, corporate bankruptcy prediction, as one of the important topics in the field of financial risk management, has made significant progress with the help of machine learning algorithms. In the future, with the continuous improvement of data collection technology and algorithm optimisation, it is believed that machine learning will play an increasingly important role in the field of corporate bankruptcy prediction and provide more reliable and efficient risk management tools for all parties. At present, the long short-term memory network (LSTM) shows great potential for application, in order to take advantage of the potential of LSTM, this paper uses LSTM to optimise the Adaboost model for corporate bankruptcy prediction, which provides a new way of thinking for corporate bankruptcy prediction.

2. Data sources

Data were collected from the Taiwan Economic Journal over a ten-year period. The definition of corporate bankruptcy is based on the commercial regulations of the Taiwan Stock Exchange. The data contains a total of 2,550 entries, each of which lists various business indicators of the firm, such as ROA(C), pre-tax interest and depreciation ROA(A), pre-tax interest and depreciation ROA(B), after-tax interest and depreciation, operating gross margin, realised sales gross margin, operating profit margin, pre-tax net interest margin, after-tax net interest margin, non-industrial income/expense/revenue, sequential interest rate (after-tax), and operating expense ratio, cash flow rate, interest-bearing debt interest rate, and tax rate, etc., and the predictor is corporate bankruptcy (1 indicates normal operations and 2 indicates bankruptcy), we display some of the data as shown in Table 1.

Table 1. Part of the dataset.

Gross Profit to Sales	Liability to Equity	Degree of Financial Leverage (DFL)	Net Income Flag	Equity to Liability	Bankrupt
0.60145329	0.290201893	0.026600631	1	0.016468741	1
0.610236526	0.28384598	0.26457682	1	0.020794306	1
0.60144934	0.290188533	0.02655472	1	0.016474114	1
0.583537612	0.281721193	0.026696634	1	0.023982332	1
0.59878151	0.27851379	0.024751848	1	0.035490201	1
0.590172327	0.2850871	0.026675366	1	0.019534478	1
0.619948867	0.292504124	0.026622298	1	0.015663075	2
0.60173934	0.278607306	0.027030517	1	0.034888556	2
0.603613451	0.276422514	0.02689063	1	0.065826497	2
0.599205074	0.279387519	0.027243015	1	0.030800865	2
0.614021193	0.278356432	0.026971091	1	0.036571691	2
0.623709203	0.277892082	0.027390858	1	0.04038102	2

3. Pearson correlation analysis

Pearson's correlation analysis is a statistical method used to measure the degree of linear correlation between two variables and is commonly used to understand the relationship between two variables as well as to predict trends between them. By calculating the covariance and standard deviation of the two variables, a Pearson's correlation coefficient ranging from -1 to 1 is finally obtained. Correlation analysis of some of the variables with corporate insolvency is carried out and correlation heat map is plotted and the results are shown in Figure 1.

DOI: 10.54254/2755-2721/87/20241631

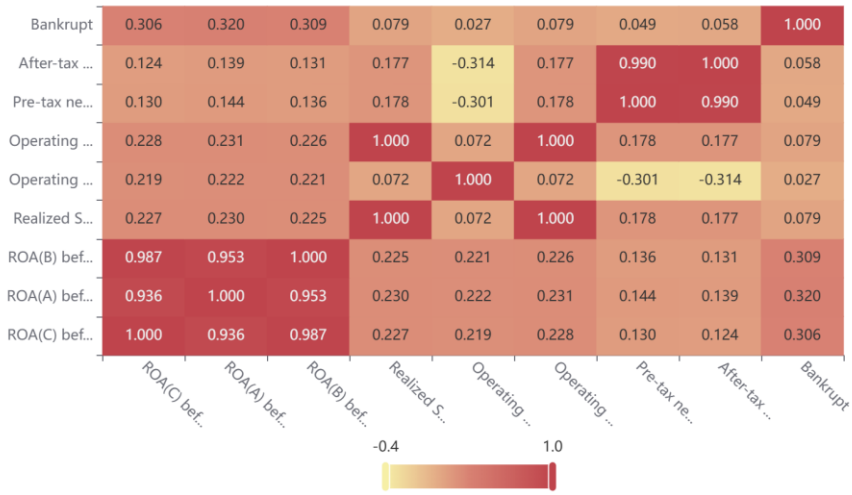

Figure 1. Correlation heat map.

From the correlation heat map, it can be seen that there is a relatively strong correlation relationship between ROA(C), pre-tax interest and depreciation ROA(A), pre-tax interest and depreciation ROA(B), after-tax interest and depreciation, operating gross margin and realised gross sales margin and corporate bankruptcy, and it is possible to use the machine learning method to predict corporate bankruptcy.

4. Pearson correlation analysis

4.1. Long Short-Term Memory

Long Short-Term Memory (LSTM) network is a deep learning model commonly used to process sequential data, LSTM network solves the long-term dependency problem in traditional Recurrent Neural Networks (RNN) by introducing a gating mechanism.

The core principle of LSTM is that its internal structure contains three key gating units: forgetting gate, input gate and output gate. These gating units help the LSTM network to learn long-term dependencies and efficiently capture important information in sequential data [7].The network structure of LSTM is shown in Fig. 2.

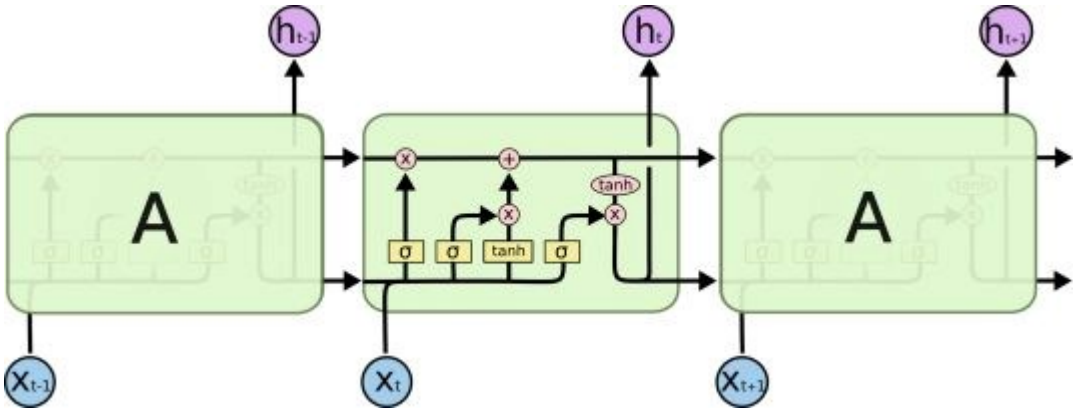

Figure 2. The network structure of LSTM.

Firstly, there is the forgetting gate, which determines which information in the past memory needs to be retained and which information needs to be forgotten. A value between 0 and 1 is output through

DOI: 10.54254/2755-2721/87/20241631

a sigmoid activation function indicating how much information needs to be retained in the corresponding position in each memory cell.

Next comes the input gate, which is responsible for updating the contents of the memory cells. It first determines which information needs to be updated through the sigmoid activation function, and then generates a new candidate value to be used to update the memory cell through the tanh activation function.

Finally, there is the output gate, which calculates the final output based on the current input and the previously saved memory states. The sigmoid function is first used to determine which information will be output to the next layer or as the final prediction, and then the tanh function is used to map the content of the memory cells between -1 and 1 as the output [8].

4.2. Adaboost

Adaboost is an integrated learning method that aims to build a strong classifier by combining multiple weak classifiers. The principle is based on continuously adjusting the weights of the training samples so that the samples that were misclassified by the previous round of classifiers receive more attention in the next round, thus improving the accuracy of the overall model.

First, Adaboost assigns an initial weight to each training sample, which is usually equal. Then, in each round of training, Adaboost selects a weak classifier that is currently optimal and adjusts the weights of the samples based on its classification results. Samples that are misclassified will receive higher weights, while those that are correctly classified will receive lower weights. Doing so ensures that the next round of training, the samples that were misclassified in the previous round receive more attention in order to train a more accurate model [9].

Next, after each weak classifier is trained, Adaboost determines the weight that the classifier will have in the final model based on its accuracy. Classifiers with higher accuracy will be given greater weights and therefore play a greater role in the final model. By iterating this process, Adaboost is able to combine multiple weak classifiers into a powerful integrated model with good generalisation to all types of data.

4.3. LSTM Optimisation of Adaboost Algorithm

LSTM is a special kind of recurrent neural network, mainly used for processing and predicting time series data. And Adaboost is an integrated learning method that builds a more powerful classifier by combining multiple weak classifiers. Combining LSTM with Adaboost can improve the modelling and prediction of time series data by taking advantage of the long-term memory of LSTM and the integration advantage of Adaboost.

Firstly, LSTM, as an RNN, can effectively capture long-term dependencies in time series data. It controls the input, output and forgetting of information through a gating mechanism, thus retaining important information and discarding irrelevant information during training. This enables LSTM to better capture complex dependencies between data when dealing with time series data, thus improving the accuracy and generalisation of the model [10].

Second, Adaboost, as an integrated learning method, iteratively trains multiple weak classifiers and adjusts the sample weights according to their performance, eventually combining these weak classifiers into a more powerful classifier. When combining LSTM with Adaboost, LSTM can be used as the base classifier and the weights between different base classifiers can be adjusted by the Adaboost algorithm to further improve the performance of the overall model.

Firstly, LSTM is used to model and predict the time series data; then the Adaboost algorithm is trained based on the prediction results generated by the LSTM model as well as the real labels; finally, multiple LSTM models after adjusting the weights are combined into a more powerful integrated model. In this way, the respective advantages of LSTM and Adaboost can be fully exploited and better results can be achieved in time series data modelling and prediction tasks.

DOI: 10.54254/2755-2721/87/20241631

5. Result

The dataset is divided according to the ratio of 7:3, with 70% of the data used for model training and 30% of the data used to test the prediction effect of the trained model. This experiment is carried out using a local server and in the experimental setup, the maximum number of training sessions is set to 1000, the initial learning rate is set to 0.01, the learning rate reduction factor is set to 0.1, the number of hidden layer nodes is set to 6, and the number of weak regressors is set to 10. In addition, the machine used for this experiment has a CPU of 32G, a graphics card of 3090, and the experiment is based on Matlab R2019b.

In this paper, the model prediction effectiveness is evaluated using the confusion matrix as well as ACCURACY, which is a two-dimensional table that shows the accuracy of the prediction results of the classification model. Accuracy is the ratio of the number of samples correctly predicted by the model to the total number of samples, and it is one of the most intuitive assessment metrics. The confusion matrix for the training set and test set is shown in Figure 3.

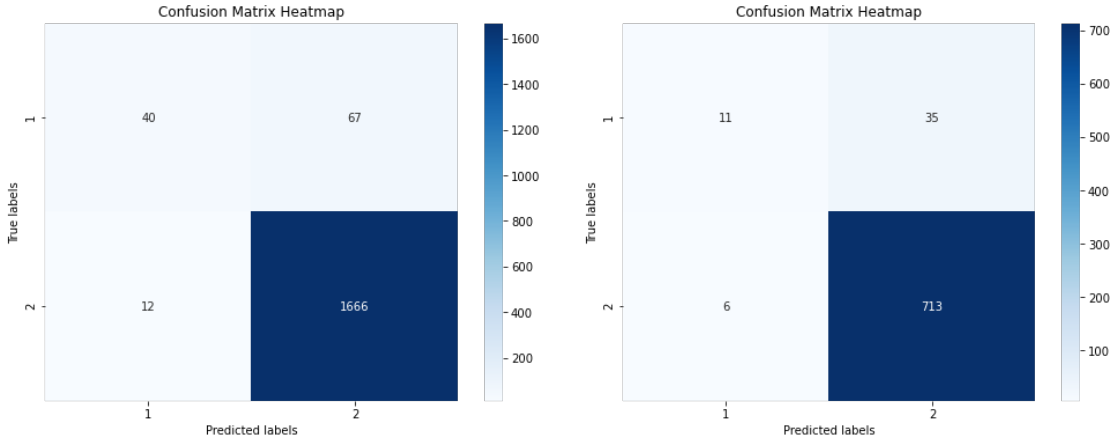

Figure 3. The confusion matrix for the training set and test set.

From the confusion matrix, it can be seen that there are 1706 correct predictions and 79 incorrect predictions in the training set, in which 12 enterprises that should have been predicted as bankrupt are predicted as normal operation and 67 enterprises that should have been predicted as normal operation are predicted as bankrupt. There were 724 correct predictions and 41 incorrect predictions in the test set, of which 6 businesses that should have been predicted to be insolvent were predicted to be operating normally and 35 businesses that should have been predicted to be operating normally were predicted to be insolvent. The prediction accuracy of the training set is 95.57% and the prediction accuracy of the test set is 94.39%, and the model achieves a good prediction effect and generalisation ability.

6. Conclusion

In this paper, we use Long Short-Term Memory Network (LSTM) to optimise the Adaboost model to predict corporate bankruptcy, and achieve satisfactory results. Firstly, this paper uses LSTM to model and predict time series data, making full use of the advantages of LSTM network in processing series data. Subsequently, the Adaboost algorithm was trained by combining the prediction results of the LSTM model with real labels to further improve the accuracy and stability of the model. Finally, multiple LSTM models with adjusted weights were combined into a more powerful integrated model to further enhance the prediction results.

The results of the confusion matrix analysis show that 1,706 samples were correctly predicted in the training set, and only 79 samples were incorrectly predicted. Among them, 12 enterprises that should have been predicted as bankrupt were misjudged as normal operation, and 67 enterprises that should have been predicted as normal operation were misjudged as bankrupt. A total of 724 samples in the test

DOI: 10.54254/2755-2721/87/20241631

set were correctly predicted, and only 41 samples were incorrectly predicted. Specifically, in the test set 6 firms that should have been predicted as insolvent were mispredicted as operating normally and 35 firms that should have been predicted as operating normally were mispredicted as insolvent.

Overall, an accuracy of 95.57% and 94.39% was achieved on the training and test sets, respectively. This indicates that the proposed LSTM-based optimised Adaboost model for predicting corporate bankruptcy has high accuracy and generalisation ability. This method not only performs well on the training set, but also can achieve quite good results on unknown data. By combining LSTM with Adaboost and constructing an integrated model using multiple LSTM models, accurate prediction of corporate bankruptcy can be effectively performed. This method not only improves the prediction effect and generalisation ability, but also provides a new idea and method for enterprise risk management. It is hoped that this method can play a greater role in practical applications and promote more in-depth exploration and application in related fields.

References

[1] Šneidere, Ruta, and Inga Būmane. "INSOLVENCY OF A COMPANY AND THE METHODS OF FINANCIAL ANALYSIS TO FORECAST IT." Economics & Management (2007).

[2] Antonowicz, Paweł, Kamila Migdał-Najman, and Krzysztof Najman. "Financial predictors of corporate insolvency-assessment of the forecast horizon of variables in models of early warning against corporate bankruptcy." e-mentor. Czasopismo naukowe Szkoły Głównej Handlowej w Warszawie 101.4 (2023): 39-44.

[3] Kušter, Denis, et al. "Early Insolvency Prediction as a Key for Sustainable Business Growth." Sustainability 15.21 (2023): 15304.

[4] DI CARLO, A. "Forecasting and preventing bankruptcy: A conceptual review." African journal of business management 12.9 (2018): 231-242.

[5] Voda, Alina Daniela, et al. "Corporate bankruptcy and insolvency prediction model." Technological and Economic Development of Economy 27.5 (2021): 1039-1056.

[6] Correia, Cláudia, et al. "How can insolvency in tourism be predicted? The case of local accommodation." International Journal of Tourism Cities 8.4 (2022): 1127-1140.

[7] Alexandrovna-Chernyavskaya, Svetlana, et al. "Practical means to forecast potential bankruptcy and financial insolvency of companies." Revista de Investigaciones Universidad del Quindío 34.S2 (2022): 276-283.

[8] Aleksandrovna, Borisuk Anastasia. "Analysis and forecasting of financial insolvency of enterprises." (2023).

[9] CHERNYAVSKAYA, SVETLANA ALEXANDROVNA, et al. "Analytical tools for forecasting financial insolvency and potential bankruptcy of a company." The journal of contemporary issues in business and government 27.2 (2021): 3937-3943.

[10] Dunis, Christian L., and J. Alexandros Triantafyllidis. "Alternative forecasting techniques for predicting company insolvencies: The UK example (1980-2001)." Neural Network World 13.4 (2003): 326-360.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241570

The application of machine learning in the field of biomedical science

Zijing Li

The Department of Internet of Things Engineering, Changsha University, Changsha, 410022, China

lizijing@ldy.edu.rs

Abstract. The application of Machine Learning (ML) in the field of biomedical science has been rapidly evolving, providing novel solutions for complex challenges such as disease prediction, drug design, and personalized medicine. This paper presents an overview of ML applications in biomedicine, focusing on three main areas: cancer prediction, common disease prediction, and drug design. The introduction to machine learning workflow is briefly discussed, highlighting the essential steps involved in training ML models with biomedical data. In the realm of cancer prediction, ML algorithms are used to analyze large datasets containing patient information, genetic profiles, and clinical variables to predict cancer susceptibility and progression. Similarly, in predicting common diseases, ML techniques are employed to identify patterns and risk factors associated with conditions like diabetes, cardiovascular diseases, and respiratory illnesses. Furthermore, ML plays a critical role in drug design by accelerating the discovery process, optimizing pharmacokinetic properties, and predicting drug responses. The discussion section delves into the potential impacts and limitations of ML in these areas, emphasizing the need for accurate data, robust validation methods, and interdisciplinary collaborations. In conclusion, the integration of ML in biomedical science offers transformative opportunities for improving healthcare outcomes. However, careful consideration of ethical, legal, and social implications is necessary as these technologies continue to advance.

Keywords: Machine Learning, Biomedical Science, Cancer Prediction, Disease Prediction.

1. Introduction

The application of machine learning in the biomedical field refers to the use of machine learning algorithms and techniques to analyze and interpret biomedical data, thereby uncovering biological mechanisms, assisting in disease diagnosis, predicting disease progression, optimizing treatment regimens, discovering biomarkers, and accelerating drug development. With the popularization of high-throughput technologies such as genetic sequencing, proteomics, and metabolomics, the biomedical field has generated a massive amount of data. This data contains precious information about disease mechanisms, drug actions, and individual differences. However, traditional data processing methods are unable to effectively analyze these complex datasets. Machine learning can assist in identifying new biomarkers, which can be used for early diagnosis of diseases, treatment monitoring, and prognosis assessment. Precision medicine aims to customize treatment plans based on

DOI: 10.54254/2755-2721/87/20241570

an individual's genetic information, lifestyle, and environmental factors. Machine learning is capable of processing and analyzing multidimensional data, helping doctors make more accurate diagnoses and treatment decisions. It can detect early signs of diseases from health data, which is beneficial for early intervention to prevent the onset of diseases or to reduce their severity.

Multivariate analysis and Machine Learning (ML) methods have been used to analyze these spectral datasets. Reduction and clustering are two forms of "unsupervised" machine learning. Dimensionality reduction algorithms project data into a lower dimensional (usually two or three dimensional) space to preserve as much of the original information as possible [1]. Common algorithms include principal component analysis (PCA), T-distributed Random neighborhood Embedding (tGSNE), and Unified Manifold Approximation and Projection (UMAP) [2]. The clustering algorithm clearly divides each observation into discrete groups based on their similarity to each other, which increases visualization and facilitates interpretation of high-dimensional data. Common algorithms such as K-means clustering. Distinct from the processes of dimensionality reduction and cluster analysis, predictive modeling engages in supervised learning, where it establishes a connection between the measured attribute linked to every data instance in the dataset and the corresponding target value it aims to predict. Traditional supervisory models include linear discriminant analysis (LDA), partial least-squares discriminant Analysis (PLSGDA), support vector machine (SVM), K-nearest neighbor algorithm (KNN), and decision tree-based models - random forest (RF) and regression tree (CRT) [3]. In addition, deep learning methods using complex artificial neuron structures enable advanced feature and pattern recognition. Among them, convolutional neural networks (CNNS) use shared weight filters and pooling layers in their architecture, showing higher specificity and sensitivity. When evaluating the model, in the case that the sample size is not enough to form a large number of test sets, cross-validation can evaluate the model performance by omitting a validation set during training [4], and constructing multiple permutations of the training set and validation set, such as K-fold multiple cross-validation method, leave one method, etc. The use of cross-validation must be handled carefully, selecting representative features (variables) or increasing large sample sizes based on the complexity of the features, otherwise, it will be easy to generate overfitted models, that is, high performance on the training set and poor performance on the test set or validation set. In addition to its direct application as a forecasting tool, many monitoring models can perform the function of "feature selection", in which the model ranks all the features in the input in order of importance [5], and only the most important features used to predict a particular outcome are identified and included in the final model [6]. Typically, these selected features encapsulate biological insights that correspond to potential therapeutic targets, the molecular mechanisms underlying disease, or serve as biomarkers in the context of diagnosing or tracking specific cancers.

The research objective of this article is to elucidate the extensive data analysis and processing capabilities of machine learning in the biomedical field. This article summarizes a considerable body of research and methods in the biomedical field from multiple perspectives, providing an overview of the machine data processing process and methods and introducing new approaches.

2. Method

2.1. Introduction to machine learning workflow

Machine learning, as a branch of artificial intelligence, aims to relate problems learned from data samples to general concepts of reasoning. The learning process of machine learning shown in Figure 1 consists of two stages: first, learning from a given data set to an unknown pattern. Second, use the learned patterns to predict new outputs based on the inputs. Machine learning can be primarily categorized into two types: supervised learning and unsupervised learning. In supervised learning, a dataset with labeled examples is utilized in the initial phase of the learning process, as opposed to unsupervised learning, where data without labels is examined. Supervised learning is usually divided into classification problems and regression problems. Classification problem refers to the process of

DOI: 10.54254/2755-2721/87/20241570

learning to divide input data into a finite set of classes. Regression is the problem of learning to map the input data to real values.

When applying machine learning methods, data samples form the basic components. Each sample can be described by several attributes, each consisting of a different type of value. In addition, knowing the specific type of data used in advance allows for the right choice of tools and techniques for analysis. Of course, in order to more accurately realize the effect of data analysis in machine learning, some other issues related to data should be handled well, including data quality improvement and data preprocessing steps, so that the used data is more suitable for training. Data quality challenges encompass various issues such as noise, anomalies (outliers), incomplete or missing data, and data that lacks representativeness. When the quality of the data is improved, the quality of the results is usually improved accordingly. In addition, in order to make the raw data more suitable for further analysis, the data preprocessing step should be adopted, focusing on the transformation of the data. Data preprocessing can use many different techniques and strategies with the goal of transforming the data to better fit a particular approach. Among these techniques, the most important methods include dimensionality reduction, feature extraction and feature selection.

After data preprocessing is completed, the model is trained using the data. In the process of model training, there will be training errors and testing errors. The former is the classification error of training data, and the latter is the error of test data. A good classification model should be able to fit the training set well and be able to classify all instances accurately. If the test error of the model starts to increase while the training error is decreasing, the model overfitting phenomenon will occur.

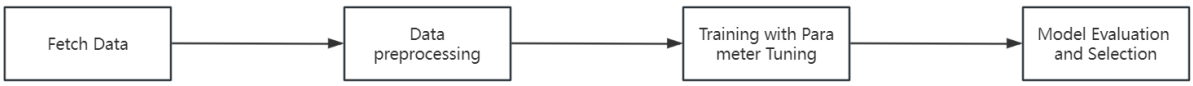

Figure 1. Machine learning workflow diagram (Photo/Picture credit: Original).

2.2. Cancer prediction

2.2.1. Prediction of lung cancer

In recent years, deep learning has made significant progress in the field of image recognition, opening up new avenues for the diagnosis and prediction of lung cancer A new predictive model of lung cancer based on convolutional neural networks (CNN) combined with attention mechanism was proposed. The model is able to automatically extract useful features from lung CT images and improve the accuracy of predictions by focusing on areas most likely to contain lesions through attention mechanisms. Arun Kumar Rana et al. discussed the evolution of machine learning in biomedical engineering, which includes applications in cancer prediction [1]. Lele Ye et al. identified potential N6-methyladenosine effector-related lncRNA biomarkers for serous ovarian carcinoma using machine learning methods [7].

2.2.2. Prediction of brain cancer

In addition to lung cancer, machine learning techniques have also been widely used in brain cancer prediction. By analyzing the patient's MRI or CT image data, combined with the patient's clinical information, researchers can build machine learning models to predict the probability of brain cancer. This predictive method can help doctors assess patients' conditions more accurately and develop more reasonable treatment plans for patients. Muhammad Shahid Shamim et al. explored the role of artificial intelligence and machine learning in medical education, including their use in understanding and predicting diseases like cancer [8].

DOI: 10.54254/2755-2721/87/20241570

2.3. Prediction of common diseases

2.3.1. Prediction of heart disease

Heart disease remains one of the leading causes of mortality worldwide, necessitating novel approaches for early detection and prevention. Machine learning has become a formidable instrument within this sphere, capable of processing extensive datasets and uncovering subtle patterns that might escape the detection of conventional statistical approaches. This section delves into the application of machine learning techniques in predicting heart disease, shedding light on the various algorithms used, the features or biomarkers identified, and the performance metrics achieved.

Machine learning algorithms utilized for heart disease prediction run the gamut from simple logistic regression models to complex deep learning architectures. One of the critical steps in developing these predictive models is feature selection, which involves choosing the most relevant biomarkers from a patient's medical history and current health profile. Characteristics commonly associated with this condition encompass factors such as age, gender, blood pressure readings, cholesterol levels, smoking habits, and a family history of heart disease, among others. These data points are subsequently input into predictive models to estimate the probability of an individual being at risk for heart disease in the future. Wu Hang et al. developed a novel causal inference algorithm for personalized biomedical causal graph learning, which can be applied to the prediction of common diseases [3].

Various studies have used these metrics to compare the performance of different machine learning models. For instance, a meta-analysis conducted by compared the efficacy of neural networks, decision trees, and logistic regression models across multiple heart disease prediction datasets. The analysis revealed that while neural networks had the highest average AUC, decision trees had the best balance between sensitivity and specificity. These comparisons are vital for determining the most suitable model for a particular clinical setting or population.

Moreover, the integration of machine learning models into clinical practice requires attention to practical considerations such as interpretability and computational efficiency. While deep learning models may offer higher accuracy, their "black box" nature can be a hindrance in healthcare settings where explainability is paramount. Conversely, models such as logistic regression, while more transparent, might not offer the sophisticated insights provided by more intricate model. Balancing these factors is crucial for the successful deployment of machine learning in heart disease prediction.

2.3.2. Stroke prediction

Stroke, much like heart disease, is a major health concern with significant morbidity and mortality rates. The ability to predict strokes before they occur could greatly improve patient outcomes by enabling early intervention. Machine learning has been at the forefront of efforts to predict stroke risk, leveraging a variety of algorithms, biomarkers, and performance metrics to achieve accurate predictions. Huan Jia Ming et al. constructed a biomedical knowledge graph of symptom phenotype in coronary artery plaque, aiding in the prediction and analysis of common cardiovascular diseases [4].

The algorithms employed for stroke prediction in machine learning are similar to those used for heart disease, encompassing techniques such as support vector machines, random forests, gradient boosting machines, and neural networks. However, the selection of features or biomarkers often differs, reflecting the unique physiological aspects of stroke. Common features include blood pressure, body mass index (BMI), history of diabetes, smoking and alcohol consumption habits, prior instances of transient ischemic attack (TIA), and the presence of carotid plaque, among others.

Performance metrics for stroke prediction are similarly focused on accuracy, sensitivity, specificity, and AUC. However, given the time-sensitive nature of stroke interventions, models are also evaluated based on their speed and computational efficiency. Rapid prediction is crucial for providing timely treatment to patients who may be experiencing a stroke. Therefore, models must not only be accurate but also capable of generating predictions quickly enough to make a difference in emergency

DOI: 10.54254/2755-2721/87/20241570

situations. Ensemble methods have likewise been explored in stroke prediction, aiming to combine the strengths of multiple models.

The interpretability issue is even more pertinent in stroke prediction due to the immediate consequences of misdiagnosis. Researchers are thus exploring techniques such as local interpretable model-agnostic explanations (LIME) and SHAP values to explain model predictions in understandable terms. These methods help clinicians understand why a model made a particular prediction, which is essential for gaining trust and promoting the adoption of machine learning in clinical practice. Jason Nan et al. used personalized machine learning-based prediction to assess wellbeing and empathy in healthcare professionals, which can be associated with their ability to predict and respond to common diseases [9].

In conclusion, machine learning has shown immense promise in predicting both heart disease and stroke, two of the most significant health concerns globally. Through a variety of algorithms, biomarkers, and performance metrics, researchers have made strides in improving prediction accuracy and efficiency. However, challenges remain, particularly in balancing model complexity with interpretability and speed. As machine learning continues to evolve, so too will its applications in predicting and preventing these life-threatening conditions.

2.4. Drug design

Drug design is one of the important research directions in biomedical science. Traditional approaches to drug design require a lot of time and resources, and have a low success rate. Machine learning techniques can build predictive models to predict the biological activity of new compounds by analyzing known data on compound structure and activity. This prediction method can greatly accelerate the process of drug design and improve the efficiency of new drug research and development. In addition, machine learning techniques can be used to optimize the structure and properties of drug molecules for better efficacy and lower side effects. Hyung Eun An et al. developed a two-layer machine learning model for the classification of legal and illegal poppy varieties, which has potential applications in drug design and development [10]. Diwei Zhou et al. emphasized the combination of data-driven machine learning approaches with prior knowledge for robust medical image processing and analysis, crucial for drug design and testing phases [11].

3. Discussion

In the modelling process, Convolutional Neural Networks (CNNs) were employed for diabetes risk prediction. Originally designed for image analysis, CNNs were adapted to handle time-series data such as fluctuations in blood sugar levels. Additionally, Random Forest and Support Vector Machines (SVMs) were integrated into the framework, with their predictions evaluated alongside CNNs. The outcomes were presented in a graphical user interface, facilitating a visual understanding of the diabetes risk. The objective was to leverage multiple models for enhanced accuracy in risk assessment through comparative analysis and visualization. Medical data is often faced with problems such as high dimension of data features, redundant features and irrelevant features. For some specific machine learning algorithms, it is unknown which features are effective for the model. In order to reduce redundant information in the data, improve the efficiency of model construction, increase the interpretability of the model and improve the model generalization performance, it is necessary to select the beneficial features of the model from all the features of the data set. Common feature extraction methods include Principal Component Analysis (PCA), mutual information, etc. Through the visual analysis of the diabetes data set in this paper, it is known that the cause of diabetes is determined by multiple factors, and the mutual information method only examines the influence of a single feature on the target variable when selecting features. Thus, effective features cannot be extracted; PCA method is used for dimensionality reduction, but not all the data in diabetes data set conform to normal distribution, so the extracted principal components are not optimal. Moreover, the data in this dataset has many features with small variance, and the correlation between the features is not taken into account, so PCA and mutual information method cannot satisfy the selection of data

DOI: 10.54254/2755-2721/87/20241570

features in this dataset. Since XGBoost algorithm has the function of helping researchers to clarify the influence degree of a specific feature on the label, and has excellent classification effect and robustness on data sets with small sample size, recursive feature elimination method based on XGBoost feature importance, namely RFE-XGBoost, is used in feature selection. The optimal feature subset is selected.

In the health management of diabetic patients, patients upload and store health indicators through a prediction system to help individuals adjust their health status and lifestyle. Diabetes prediction is based on personal health data. With the development of big data technology, data collection is more convenient and the accuracy is greatly improved. Intelligent blood glucose meter, electronic scale, health bracelet and other devices can dynamically obtain health data, and diabetes prediction model can be used to achieve prediction.

First, all the features of the pre-processed data set are input into the model, and the XGBoost algorithm model is used to sort the feature importance, and the whole data set is fitted to complete the preliminary feature screening. When XGBoost algorithm is used to sort the features of the pre-processed data, GridSearchCV (grid search method) is used to find the optimal parameters of the model. First, the learning_rate, n_estimators and max_depth of the tree, three main parameters that determine the model performance, are put into the dictionary param_grid variable as keys. The value of the key is set by arange() in the numpy module, and then the required parameters are put into the Grid Search CV() function, where the value of the parameter cv is set to 10. Then best_params, best_score_, best_estimator_ and best_index_ are used to output the value of the optimal parameter, the score of the model under the optimal parameter, the model under the optimal parameter and the index under the optimal parameter. Finally, the feature importance of the data set is obtained. The plot_importance module in XGBoost library is used to check the feature importance and order.

The XGBoost Special importance score is determined by the sum of the number of splits per tree for a particular feature. For example, if the feature splits once in the first tree, twice in the second tree, three times in the third tree, and so on, then the score of the feature is the sum of the number of times the feature splits on all trees. According to the ranking of feature importance in XGBoost model, the features with zero scores are eliminated in this paper, and the number of remaining features after screening is 20. In the way of recursive elimination, the first n features of feature importance ranking are selected each time to form a feature subset. According to the evaluation index AUC performance of the classifier, the best feature subset of the n feature subsets is selected at last.

Therefore, the optimal feature subsets of the diabetes data set in this paper were HBA1c, blood glucose value, age, BMI, waist circumference, triglyceride, low-density lipoprotein, urea, uric acid, total cholesterol, and alanine aminotransferase. It can be seen from the features in the selected optimal feature subset that both glycosylated hemoglobin and blood sugar are the most favorable indicators for judging diabetes, which is consistent with the medical rules. However, in the traditional diagnosis process, doctors' personal experience and subjective judgment are often relied on, while the feature selection in this paper ranks the importance of diabetes features. In addition, the selection of diabetes features was quantified, and different features had different degrees of impact on the contribution of the prediction model, so that the risk factors inducing diabetes could be known more scientifically and accurately.

4. Conclusion

In conclusion, the applications of machine learning in the biomedical field have revolutionized the way people approach healthcare, diagnosis, and drug discovery. The capacity of machine learning algorithms to handle extensive datasets and reveal patterns invisible to the human eye has pioneered new territories in precision medicine and personalized therapies. From improving the accuracy of disease diagnosis to optimizing treatment plans, machine learning techniques are poised to transform healthcare delivery and patient outcomes.

Moreover, as the technology continues to evolve and become more sophisticated, scientists expect to see even more innovative applications in the biomedical field. Machine learning's potential in

DOI: 10.54254/2755-2721/87/20241570

predictive modelling, risk assessment, and biomarker discovery is particularly promising, as it holds the key to earlier intervention, prevention, and ultimately, the eradication of many chronic and life-threatening diseases.

However, it is also crucial to recognize the challenges that accompany the use of machine learning in biomedical research, including data privacy, interpretability, and ethical considerations. Moving forward, it is crucial to tackle these challenges ethically and sustainably to maximize the benefits of machine learning while mitigating potential risks.

References

[1] Rana AK, Sharma V, Rana SK & Chaudhary VS 2024 Evolution of Machine Learning and Internet of Things Applications in Biomedical Engineering CRC Press: 2024-06-13

[2] Goel N & Yadav RK 2024 Internet of Things enabled Machine Learning for Biomedical Application CRC Press: 2024-03-30

[3] Wu H, Shi W & Wang MD 2024 Developing a novel causal inference algorithm for personalized biomedical causal graph learning using meta machine learning BMC Medical Informatics and Decision Making vol 24 (1) pp 137-137

[4] Huan JM, Wang XJ, Li Y, Zhang SJ, Hu YL & Li YL 2024 The biomedical knowledge graph of symptom phenotype in coronary artery plaque: machine learning-based analysis of real-world clinical data BioData Mining vol 17 (1) pp 13-13

[5] Prasad A, Santra TS & Jayaganthan R 2024 A Study on Prediction of Size and Morphology of Ag Nanoparticles Using Machine Learning Models for Biomedical Applications Metals vol 14 (5)

[6] Labory J, Fotso EN & Bottini S 2024 Benchmarking feature selection and feature extraction methods to improve the performances of machine-learning algorithms for patient classification using metabolomics biomedical data Computational and Structural Biotechnology Journal vol 23 pp 1274-1287

[7] Ye L, Tong X, Pan K, Shi X, Xu B, Yao X, Zhuo L, Fang S, Tang S, Jiang Z, Xue X, Lu W & Guo G 2024 Identification of potential novel N6-methyladenosine effector-related lncRNA biomarkers for serous ovarian carcinoma: a machine learning-based exploration in the framework of 3P medicine Frontiers in Pharmacology vol 15

[8] Shamim MS, Zaidi SJA & Rehman A 2024 The Revival of Essay-Type Questions in Medical Education: Harnessing Artificial Intelligence and Machine Learning Journal of the College of Physicians and Surgeons--Pakistan : JCPSP vol 34 (5) pp 595-599

[9] Nan J, Herbert MS, Purpura S, Henneken AN, Ramanathan D & Mishra J 2024 Personalized Machine Learning-Based Prediction of Wellbeing and Empathy in Healthcare Professionals Sensors (Basel, Switzerland) vol 24 (8) pp 2640-2642

[10] An HE, Mun MH, Malik A & Kim CB 2024 Development of a two-layer machine learning model for the forensic application of legal and illegal poppy classification based on sequence data Forensic Science International: Genetics vol 71 p 103061

[11] Zhou D, Duan J, Qin C & Luo G 2024 Editorial: The combination of data-driven machine learning approaches and prior knowledge for robust medical image processing and analysis Frontiers in Medicine vol 11

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241620

Evaluation and optimization of intelligent recommendation system performance with cloud resource automation compatibility

Kangming Xu[1a,*], Haotian Zheng[1b], Xiaoan Zhan[2], Shuwen Zhou[3], Kaiyi Niu[4]

[1a]Computer Science and Engineering, Santa Clara University, CA, USA
[1b]Electrical & Computer Engineering, New York University, New York, NY, USA
[2]Electrical Engineering, New York University, NY, USA
[3]Computer Science, The University of New South Wales, Sydney, Australia
[4]Artificial intelligence, Royal Holloway University of London, Egham, UK

[*]kangmingxu87@gmail.com

Abstract. This paper comprehensively explores the integration of cloud computing and advanced recommendation systems, emphasizing their pivotal roles in enhancing user experiences and operational efficiencies across digital platforms. It reviews the evolution of recommendation algorithms, highlighting their application in diverse domains such as e-commerce and media. The study evaluates the performance of advanced models like UniLLMRec against traditional counterparts using datasets from news and e-commerce domains. Additionally, the paper discusses the infrastructure architecture of cloud computing, demonstrating its capability to support scalable and efficient data processing. Through experimental insights and methodology, the research underscores the transformative impact of cloud technologies on optimizing recommendation system performance, thereby advancing digital engagement and competitiveness.

Keywords: Cloud Computing, Recommendation Systems, Artificial Intelligence, Big Data.

1. Introduction

With the rapid advancement of digital transformation and continuous innovation in cloud technology, cloud computing has become an indispensable infrastructure across business, government, and personal services. The cloud computing market is segmented into Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service [1] (SaaS), each playing a distinct role and driving growth. IaaS, as the foundation, offers virtualized computing resources, storage, and network services, providing enterprises with a flexible and scalable platform for application deployment. Cloud platforms facilitate centralized data management and analysis, enabling precise user behavior analysis, real-time recommendation strategy adjustments, and enhancing system intelligence and user satisfaction.

This paper examines how cloud computing can enhance the performance and effectiveness of recommendation systems, thus improving user experiences and enterprise competitiveness.

DOI: 10.54254/2755-2721/87/20241620

2. Related work

2.1. Intelligent recommendation system

A recommendation system is an information filtering system that predicts a user's behavior or preference for an item. [2] Why do we need a recommendation system? Generally speaking, the recommendation system has much data information, far greater than the user's demand. The recommendation system is produced to enable the user to quickly find the information that meets the user's needs from the massive data. Therefore, the recommendation system is used for enormous data information overload, and the amount of data is too small; it is not worth using the recommendation system.

In the early years, users of online shopping platforms needed to go step by step, according to the classification of goods or keyword search, to find their products in the mass of goods. In recent years, on the Double Eleven, ordinary consumers can quickly screen out the goods they want and receive recommendations for goods and live broadcasts that align with their preferences. [3] Later, the rise of AI technology, especially the information flow content recommendation and short video recommendation typical of Toutiao and Douyin, once again helped the recommendation system to improve significantly. Whether it was the product recommendation of shopping platforms, the anchor recommendation of live broadcasting platforms, or the video content recommendation of video platforms, more and more people began to sign that [4]AI knows itself better.AI recommendation systems have also become necessary for Internet companies' business. Data show that on some of the world's large online websites, even if the relevance of the recommended content is only increased by 1%, its sales will increase by billions; the AI recommendation system is undoubtedly a high-value system hidden behind many Internet applications.

2.2. Recommendation system algorithm

The core of the recommendation system is the recommendation algorithm. The recommendation algorithm can be simplified into a black box, and the input of the black box is various attributes and characteristics of the user and the candidate, including the user's age, gender, purchasing power, and the candidate's content, category, and release time. [5] The output of the black box is a list of recommendations for the user, ranked by preference.

Currently, the main recommendation algorithms include popularity-based algorithms, collaborative filtering algorithms, content-based and model-based algorithms, etc. Different recommendation algorithms have different preferences, advantages, and disadvantages. Still, they are all based on extensive data analysis to predict and recommend users and generate lists of items they may be interested in.

Based on the popularity of the algorithm, the simple version of the implementation can be sorted by the heat, such as the user's likes, comments, and forwarding amount to calculate the heat, and then according to the heat value to recommend sorting, this algorithm is relatively simple, the disadvantage is that the heat needs to be constantly optimized and improved, the need to integrate various factors continually changing, to have a good performance. Cloud computing infrastructure architecture

Cloud computing has become an infrastructure for several reasons:

1. Public cloud accelerates the integrated development of hardware and software and truly promotes the process of T service

Software and hardware integration is one of the development trends of T., the public cloud can be used as the "glue" of software and hardware; through the public cloud, the integration and integration between software and hardware becomes easy public cloud is responsible for managing all hardware resources, the software can be through the interface with the cloud, it can achieve the goal of software and hardware integration.

The industry has recognized the trend of IT as a service for many years; Saas (software as a service), PaaS(platform as a service), PaaS (infrastructure as a service), everything is a service. As software and hardware integration continues to deepen[6], T will be presented to users as services without

DOI: 10.54254/2755-2721/87/20241620

distinguishing between software and hardware services. No matter what kind of service, you need a carrier, and the public cloud is that carrier.

2. T technology rooted in the public cloud, its application and innovation must rely on the public cloud

Big data, artificial intelligence, and other new technology developments have been rooted in the public cloud; strictly speaking, if you leave the public cloud, there is no significant data and artificial intelligence. Artificial intelligence is based on extensive data development; without the network and data, artificial intelligence will no longer exist. Therefore, artificial intelligence development must also rely on the public cloud [7].

3. Public cloud is the platform of "convergence" and the interface of all "connectivity."

In the era of digital economy, integration is an inevitable trend; hardware and software should be integrated, T products and services should be integrated, industrialization and information technology should be integrated, all links of the industrial chain should be integrated, and industries should be integrated. [10] The development of all kinds of networks provides a channel for integration.

4. Infrastructure framework

The cloud computing infrastructure architecture takes distributed multi-cloud as the core, builds the "one cloud and multiple computing" converged base, relies on the unified management of heterogeneous resources and the distributed task collaboration framework, builds a new service system with AI running through it, supports the integrated carrying capacity of general computing, intelligent computing, supercomputing, and network convergence services, and ensures the availability of full-link services. The hierarchical system of traditional cloud architecture is retained in terms of overall architecture.

The cloud network resource construction emphasizes the distributed optimal layout of multiple types of resource pools. [8] Diversity is noted in the software and hardware resource layer, divided into CPU-based general computing infrastructure and intelligent computing infrastructure dominated by AI-accelerated chips such as GPU[9]. The distributed cloud platform manages multi-dimensional heterogeneous resources in a unified manner and implements efficient collaborative task scheduling. Based on infrastructure architecture, cloud service forms show a trend of generalization and intelligent development, carrying multiple business types and providing rich industrial digital capabilities.

In conclusion, the evolution of intelligent recommendation systems has revolutionized user engagement across various digital platforms, driven by sophisticated algorithms like collaborative filtering and content-based recommendations. [10] These systems have become indispensable for personalized user experiences in e-commerce, content streaming, and social media. Meanwhile, cloud computing has emerged as a robust infrastructure, offering scalable resources and efficient data processing capabilities crucial for supporting these advanced systems. As we move forward, the integration of cloud resource automation stands out as a pivotal factor in enhancing the agility and performance of recommendation systems. The next phase of our exploration will evaluate how automated cloud solutions can optimize these systems, ensuring seamless scalability, reliability, and operational efficiency.

3. Methodology

3.1. Experimental design

1. Data set

The experiment used two datasets related to detailed information about news recommendations and film and television reviews, respectively - MIND and Amazon Review. The former contains News articles and user behavior logs from the Microsoft News website; The latter is collected from Amazon's e-commerce platform and includes user reviews, ratings, and product information.

2. Evaluate indicators

Regarding evaluation indicators, we mainly focus on the performance of Recall and Re-ranking tasks. Indicators such as Recall, Normalized Discounted Cumulative Gain (NDCG) [11], and Intra-List Average Distance (ILAD) were used to evaluate the model's performance in the recommendation task.

DOI: 10.54254/2755-2721/87/20241620

3.2. Model comparison

The UniLLMRec framework is also compared to a series of baseline models; These include Popularity-based recommendation (Pop), Factorization Machines (FM), Deep FM, NRMS, SASRec, and LLM-Ranker.

Pop: Rank an item based on its overall user popularity in the user base and recommend the most popular item to the user.

Pros: Simple, easy to implement, and effective for widely popular content.

Disadvantages: Lack of personalization, consideration of user-specific preferences, and inability to satisfy users with specific tastes.

FM: The ability to use factorization parameters to estimate interactions between variables and can handle problems with high sparsity. Because it can combine auxiliary information to overcome the difficulties of cold start and sparse data in a recommendation system, it is a very practical recommendation model

Deep FM: Combines a shallow factorization model with a deep neural network, leveraging both strengths to improve the model's predictive power. This makes it excellent at dealing with complex interactions and sparsity in recommendation systems, especially in scenarios like [12]CTR prediction.

Advantages: The interaction between features can be learned automatically without manually designing the feature interaction. In addition, due to its profound learning nature, Deep FM scales well to large-scale data sets and can adapt to changing data distributions.

NRMS uses a multi-head self-attention mechanism to enhance the performance of the news recommendation system. The model is divided into two main parts: news encoder and user encoder. News encoders use multi-head self-attention to learn the representation of words in news articles. In contrast, user encoders utilize the exact mechanism to capture behavioral patterns and preferences in a user's reading history. In this way, the model can understand and match user interests and relevant news content more accurately.

SASRec is a sequence-based recommendation system that employs a self-attention mechanism to assign weights to past items dynamically in each time step. Its adaptive nature prioritizes long-term dependencies in dense data sets and focuses on recent activity in sparse data sets, contributing to its superior performance.

LLM-Ranker takes advantage of the rich semantic and contextual understanding of LLMS, such as GPT or BERT, to improve ranking tasks in search and recommendation systems, which makes the model more efficient at handling complex user queries and diverse content.

3.3. Performance comparison

From the direct comparison of the performance of UniLLMRec and traditional recommendation model in MIND and Amazon Review data sets and two indicators, respectively, whether GPT-3.5 or GPT-4 is used as a backbone, UniLLMRec can outperform many traditional models (especially the one with GPT-4 as the backbone) when the proportion of training sets is small, which indicates the advantage of UniLLMRec's zero sample learning cost, and also reflects the effectiveness and relevance of its retrieval and recommendation items.

DOI: 10.54254/2755-2721/87/20241620

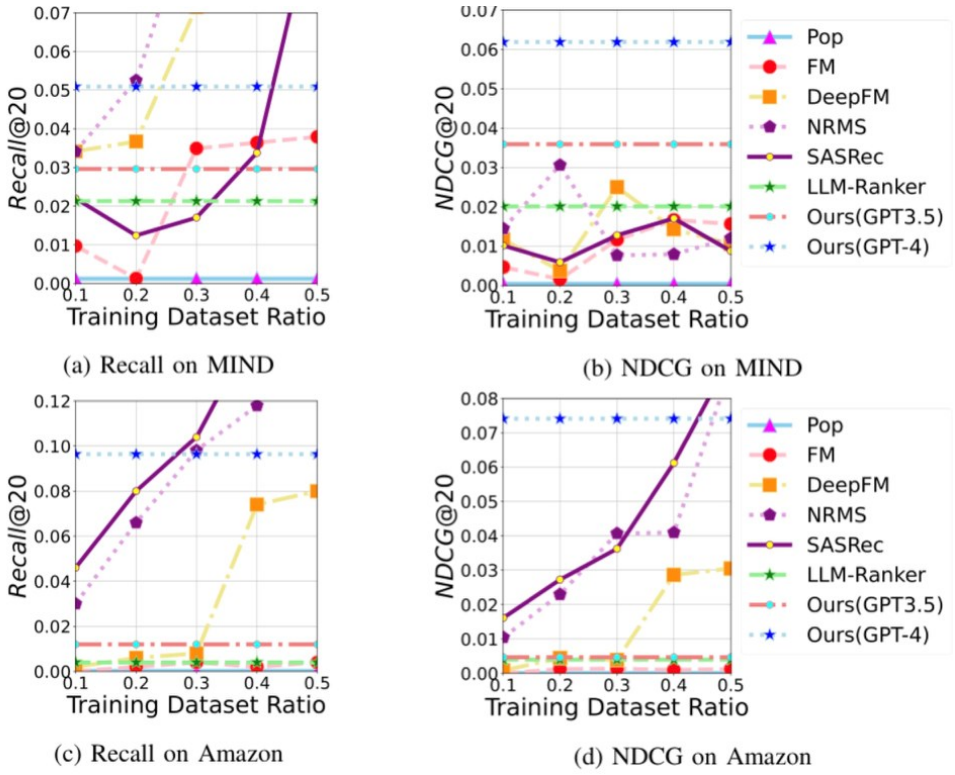

Figure 1. Performance Comparison of Recall and NDCG value on MIND and Amazon datasets.

3.4. Cloud computing deployment

When cloud computing is introduced into the experimental design of a recommendation system, the performance and scalability of the system can be significantly improved. By leveraging the elastic computing resources and efficient storage services provided by cloud computing platforms, we can more efficiently process large-scale data sets, such as MIND and Amazon Review data. During the experiment, the elastic resource characteristics of the cloud computing platform can also automatically adjust the computing resources according to the real-time load, ensuring that the recommendation system can maintain stable performance at peak times. To sum up, the introduction of cloud computing technology can not only optimize the operational efficiency and performance of the recommendation system but also improve the security and reliability of the system, providing a broader and controllable platform for the research and experiment of recommendation algorithms.

3.5. Experimental conclusion

Although UniLLMRec can complete the entire recommendation process in most cases, the researchers also found some problems during the experiment, such as even if the output format is clearly defined in the prompt template, LLM sometimes does not output items according to the instructions, resulting in the items not being correctly indexed (intention recognition problem); In the user interest modeling stage, LLM can capture and summarize user interest, but in the leaf node retrieval and diversity perception rearrangement stage, there is a risk of including examples in the wrong prompt words into the retrieval process (illusion problem).

The researchers made an interesting observation when comparing GPT-3.5 and GPT-4 on the Amazon dataset. When GPT-3.5 and GPT-4 reach the wrong child node, GPT-3.5 will usually continue to complete the subsequent process. At the same time, GPT-4 may proactively give a hint that all candidate answers do not meet the requirements (e.g., "Based on the user's interest in UFC and combat sports," None of the Character & Series subcategories provided are relevant "). This suggests GPT-4 is more accurate than GPT-3.5 in capturing user preferences.

DOI: 10.54254/2755-2721/87/20241620

4. Conclusion

Based on the extensive exploration of cloud computing and advanced recommendation systems in this study, it is evident that integrating cloud infrastructure significantly enhances the performance and scalability of recommendation systems across digital platforms. The deployment of cloud resources revolutionizes the processing of large-scale datasets, particularly in dynamic domains like news and e-commerce, where data volumes are immense and constantly evolving. Cloud infrastructure offers scalable computing power and storage capabilities that traditional on-premises systems struggle to match. This scalability enhances the speed and efficiency of data processing and ensures that recommendation systems can handle peak loads without compromising performance.

In evaluating advanced models such as UniLLMRec against traditional methods, significant advantages emerge regarding recall and recommendation accuracy. UniLLMRec leverages state-of-the-art AI technologies like large language models (LLMs) to analyze user behaviours and preferences more effectively. These models excel in understanding nuanced patterns in user interactions, delivering more personalized recommendations that align closely with individual interests and needs.

Furthermore, the study highlights the critical role of cloud computing as a foundational infrastructure supporting AI-driven recommendation algorithms. By leveraging elastic computing resources and efficient storage services, cloud platforms enable dynamic adjustments to real-time loads, ensuring consistent system performance during peak usage. This research underscores the strategic importance of cloud-enabled recommendation systems in driving digital engagement and competitiveness across various industries, paving the way for future innovations in user-centric technologies.

References

[1] Hoque, M.S., Mukit, M.A. and Bikas, M.A.N., 2012. An implementation of an intrusion detection system using a genetic algorithm. arXiv preprint arXiv:1204.1336.

[2] Bace, R.G. and Mell, P., 2001. Intrusion detection systems.

[3] Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J. and Ahmad, F., 2021. Network intrusion detection system: A systematic study of machine learning and deep learning approaches. Transactions on Emerging Telecommunications Technologies, 32(1), p.e4150.

[4] Zhan, T., Shi, C., Shi, Y., Li, H., & Lin, Y. (2024). Optimization Techniques for Sentiment Analysis Based on LLM (GPT-3)—arXiv preprint arXiv:2405.09770.

[5] Li, Huixiang, et al. "AI Face Recognition and Processing Technology Based on GPU Computing." Journal of Theory and Practice of Engineering Science 4.05 (2024): 9-16.

[6] Yuan, J., Lin, Y., Shi, Y., Yang, T., & Li, A. (2024). Applications of Artificial Intelligence Generative Adversarial Techniques in the Financial Sector. *Academic Journal of Sociology and Management*, *2*(3), 59-66.

[7] Guo, L., Li, Z., Qian, K., Ding, W., & Chen, Z. (2024). Bank Credit Risk Early Warning Model Based on Machine Learning Decision Trees. *Journal of Economic Theory and Business Management*, *1*(3), 24-30.

[8] Li, Zihan, et al. "Robot Navigation and Map Construction Based on SLAM Technology." (2024).

[9] Fan, C., Ding, W., Qian, K., Tan, H., & Li, Z. (2024). Cueing Flight Object Trajectory and Safety Prediction Based on SLAM Technology. Journal of Theory and Practice of Engineering Science, 4(05), 1-8.

[10] Ding, W., Tan, H., Zhou, H., Li, Z., & Fan, C. Immediate Traffic Flow Monitoring and Management Based on Multimodal Data in Cloud Computing.

[11] Lin, Y., Li, A., Li, H., Shi, Y., & Zhan, X. (2024). GPU-Optimized Image Processing and Generation Based on Deep Learning and Computer Vision. *Journal of Artificial Intelligence General Science (JAIGS) ISSN: 3006-4023*, *5*(1), 39-49.

[12] Yang, Z., Li, L., Lin, K., Wang, J., Lin, C. C., Liu, Z., & Wang, L. (2023). The dawn of lmms: Preliminary explorations with gpt-4v (ision). arXiv preprint arXiv:2309.17421, 9(1), 1.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241621

A comprehensive review on the application of CVSS 4.0 and deep learning in vulnerability

Hongyu Xie

Beijing University of Posts and Telecommunications, 10 Xitucheng Road, Haidian District, Beijing, China

buptxhy@163.com

Abstract. This paper reviews the evolution of the Common Vulnerability Scoring System (CVSS), focusing on the enhancements and applications of CVSS 4.0. It also explores the potential integration of deep Learning techniques in vulnerability assessment. Key findings include identifying critical improvements in CVSS 4.0 that address previous limitations and enhance accuracy and granularity in vulnerability scoring. Additionally, the paper demonstrates how deep Learning models can predict vulnerability scores and trends, thereby improving the speed and precision of assessments. By combining CVSS 4.0 with deep Learning technologies, this paper proposes a more comprehensive and efficient approach to vulnerability assessment, which could significantly enhance proactive security measures and risk management strategies.

Keywords: CVSS 4.0, Deep Learning, Vulnerability Assessment, Risk Management.

1. Introduction

With the rapid advancement of information technology, the number of security vulnerabilities in systems and applications is increasing. Effectively assessing and managing these vulnerabilities has become a crucial task in information security. The Common Vulnerability Scoring System (CVSS) provides a standardized method for scoring vulnerabilities and has been widely adopted. This paper presents a detailed introduction to the evolution of CVSS, highlighting the specific enhancements introduced in CVSS 4.0.

CVSS 4.0 has significantly improved over its predecessor, CVSS 3.1, by expanding metric groups, refining attack vectors, and distinguishing attack complexities and requirements. For instance, it introduced a new Threat metric group, provided more granular attack vector definitions, and updated user interaction metrics. Despite these advancements, existing methods still face limitations in accurately predicting and assessing vulnerabilities in dynamic and complex environments.

This paper also explores the potential applications of deep learning in vulnerability assessment. Mentioning the limitations of existing methods, such as their reliance on static scoring models and inability to adapt to evolving threats, we propose the integration of deep learning techniques. Deep learning models can dynamically analyze vast amounts of data, identify patterns, and predict vulnerability scores and trends with higher accuracy and speed. By combining CVSS 4.0 with deep learning technologies, this paper proposes a more comprehensive and efficient approach to vulnerability assessment, addressing existing gaps and significantly enhancing proactive security measures and risk management strategies.

DOI: 10.54254/2755-2721/87/20241621

2. Evolution of CVSS

2.1. CVSS 2.0

CVSS 2.0, released in 2005, includes three metric groups: Base Metrics, Temporal Metrics, and Environmental Metrics. Base Metrics assess the intrinsic characteristics of a vulnerability, Temporal Metrics consider changes in these characteristics over time, and Environmental Metrics adjust the risk assessment based on the user's environment. However, CVSS 2.0 had limitations in its granularity and adaptability, often resulting in inconsistent vulnerability assessments across different environments[1].

2.2. CVSS 3.0

CVSS 3.0, released in 2015, introduced significant improvements. New metrics such as Scope were added, and existing metrics like Attack Vector, Attack Complexity, and User Interaction were refined. These changes allowed CVSS 3.0 to more accurately reflect the actual impact of vulnerabilities[2]. The inclusion of Temporal and Environmental metrics aimed to provide a holistic view of vulnerabilities, accounting for factors like exploit availability and the impact on different industries or user populations.

2.3. CVSS 4.0

CVSS 4.0, the latest version, further enhances the scoring methodology and metrics, adding new assessment dimensions such as Vulnerability Chaining. This version emphasizes the specific impacts of vulnerabilities in different environments, providing users with a more detailed and comprehensive risk assessment tool. CVSS 4.0 introduces a fifth metric group, Scope, which distinguishes between vulnerabilities with internal and external scopes. It also includes Environmental Metrics, enabling organizations to tailor vulnerability assessments to their specific contexts, and enhancing the relevance and accuracy of vulnerability assessments across diverse organizational contexts.

3. Application of Deep Learning in Vulnerability Assessment

3.1. Overview of Deep Learning Technologies

Deep learning, a branch of machine learning, employs advanced algorithms to discern patterns within extensive datasets, facilitating predictive analytics and decision-making processes. This technology has achieved notable success across diverse domains such as image recognition, natural language processing, and data mining. In the realm of vulnerability assessment, deep learning techniques can scrutinize historical vulnerability data to identify trends, foresee potential threats, and enhance the efficiency and precision of assessments.

3.2. Applications in the Security Domain

Within the field of information security, deep learning is utilized for a myriad of purposes, including intrusion detection, malware classification, and anomaly detection. The capability of deep learning models to process and analyze extensive historical data enables them to more effectively identify potential security threats and vulnerability exploitation behaviors[3].

3.2.1. Specific Examples and Case Studies

(1) Google Project Zero: Automated Vulnerability Detection

Case Background: Google Project Zero, a security research team, is dedicated to discovering and reporting software vulnerabilities. The team has effectively utilized deep learning models for vulnerability detection.

Deep Learning Application: Project Zero employed deep learning models to analyze both binary and source code. These models identified vulnerabilities through pattern recognition and anomaly detection techniques, leveraging a large dataset of known vulnerabilities to predict similar issues in new codebases.

DOI: 10.54254/2755-2721/87/20241621

Impact: The application of deep learning significantly enhanced the efficiency of vulnerability detection, reducing the time required for manual code reviews and accelerating the identification of potential security risks.

(2) DARPA Cyber Grand Challenge: Automated Vulnerability Assessment

Case Background: The DARPA Cyber Grand Challenge aimed to advance automated cybersecurity defenses, focusing on vulnerability discovery and patching.

Deep Learning Application: Participants in the competition used deep learning algorithms to develop systems for automatic vulnerability discovery and patch generation. These systems analyzed vulnerabilities, generated patches, and applied them autonomously using deep learning techniques.

Impact: The challenge highlighted the potential of deep learning to automate complex security tasks, providing valuable insights into the development of more effective and scalable vulnerability assessment tools.

(3) MITRE ATT&CK Framework: Threat Intelligence Analysis

Case Background: The MITRE ATT&CK Framework is a comprehensive knowledge base of adversary tactics and techniques utilized in cybersecurity incidents.

Deep Learning Application: Researchers have applied deep learning techniques to analyze data within the ATT&CK Framework, identifying patterns in attacker behavior and predicting future attack vectors.

Impact: Deep learning models have enhanced threat intelligence capabilities, offering more accurate predictions of potential vulnerabilities and informing defensive strategies.

3.3. Advantages of Combining CVSS with Deep Learning

Combining the Common Vulnerability Scoring System (CVSS) with deep learning technology offers several significant advantages:

Automated Vulnerability Scoring: Deep learning models can automate the assessment of vulnerability severity, minimizing the need for manual intervention and enabling swifter response times[4].

Real-time Updates: As new vulnerabilities are discovered, deep learning models can be updated promptly, providing current risk assessments.

Accuracy: By leveraging historical data and considering contemporary environmental factors, deep learning models can deliver more precise risk assessment results. This enhanced accuracy aids organizations in prioritizing their response efforts more effectively.

4. Methodology

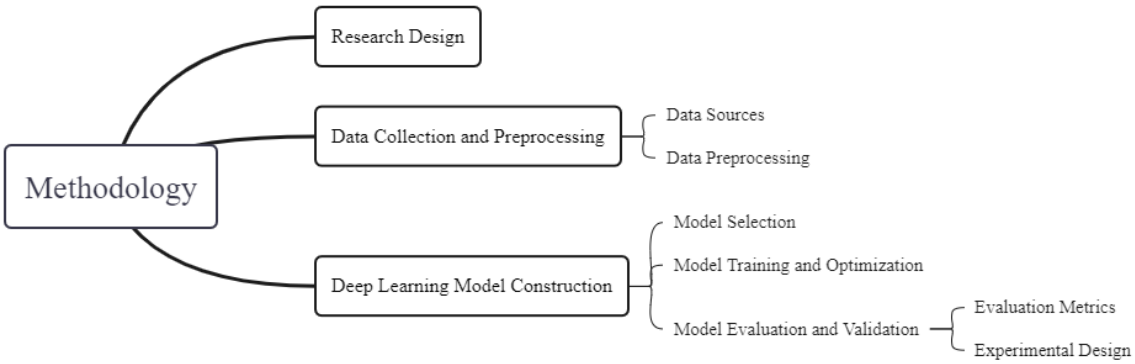

Figure 1. Methodological framework

DOI: 10.54254/2755-2721/87/20241621

4.1. Research Design

This study adopts a mixed-method approach, combining quantitative and qualitative analyses to explore how deep learning can enhance vulnerability scoring accuracy and efficiency within the CVSS 4.0 framework. The research will be conducted in several phases: data collection and preprocessing, model construction, model evaluation and validation, and results analysis and discussion.

4.2. Data Collection and Preprocessing

4.2.1. Data Sources

Data will be collected from multiple public vulnerability databases, such as the National Vulnerability Database (NVD) and the Common Vulnerabilities and Exposures (CVE). These datasets will include information on vulnerability descriptions, CVSS scores, environmental factors, temporal factors, and other relevant details.

4.2.2. Data Preprocessing

Data preprocessing will involve data cleaning, feature selection, and feature extraction. Initially, the data will be cleaned to remove missing values and outliers. Feature selection methods, such as chi-square tests and mutual information, will be used to identify the most influential features on vulnerability scoring. Feature extraction will convert textual descriptions into numerical features using techniques like the bag-of-words model and TF-IDF. To address data imbalance, methods such as the Synthetic Minority Over-sampling Technique (SMOTE), under-sampling, and class weight adjustments will be employed.

4.3. Deep Learning Model Construction

4.3.1. Model Selection

Based on existing research and data characteristics, several common learning algorithms will be selected, including Support Vector Machines (SVM), Random Forests (RF), and Gradient Boosting Decision Trees (GBDT). SVM performs well in high-dimensional spaces and is suitable for complex classification tasks; RF has good generalization ability and can effectively handle high-dimensional data; GBDT excels in capturing non-linear relationships and feature interactions.

4.3.2. Model Training and Optimization

Models will be trained and optimized using techniques such as cross-validation. Cross-validation effectively prevents overfitting and improves the model's generalization ability. During model training, emphasis will be placed on feature robustness and computational efficiency to ensure the model's applicability across different scenarios. Specific steps include splitting the data into training and validation sets, tuning model parameters using grid search and random search, and evaluating and adjusting model performance.

4.3.3. Model Evaluation and Validation

(1) Evaluation Metrics

Model performance will be comprehensively evaluated using metrics such as accuracy, precision, recall, and F1 score. Accuracy measures the overall correctness of the model's predictions; precision evaluates the model's ability to correctly identify positive samples; recall measures the proportion of actual positive samples correctly identified by the model; the F1 score balances precision and recall. Additionally, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and Precision-Recall (PR) curve will be used to assess model performance on imbalanced data.

DOI: 10.54254/2755-2721/87/20241621

(2) Experimental Design

Multiple experiments will be designed to verify the model's applicability and stability across different datasets and environmental complexities. Specific experiments include training and testing the model on different datasets, evaluating model performance under varying environmental factors (such as operating systems and software types), and assessing performance over different periods (such as new vs. old vulnerability data). Comparative analysis will identify the optimal model and its applicable scenarios.

5. Results Analysis and Discussion

Once the experimental results are obtained, it is crucial to structure the discussion to directly address the research questions and objectives stated in the introduction. This section explores the effectiveness of deep learning models within the CVSS 4.0 framework, evaluates the advantages and limitations of CVSS 4.0, and identifies future research directions.

5.1. CVSS 4.0 Advantages and Limitations

(1) Advantages:

CVSS 4.0 introduces significant advancements over its predecessors, enhancing the granularity and relevance of vulnerability assessments. The introduction of the new Threat metric group, refined attack vector definitions, and the Scope metric provides a more detailed and nuanced approach to evaluating vulnerabilities. Our results demonstrate that CVSS 4.0 offers improved accuracy in scoring vulnerabilities across diverse organizational contexts compared to CVSS 3.1. For instance, the detailed Scope metric enabled more precise differentiation between internal and external vulnerabilities, leading to more accurate risk assessments.

(2) Limitations:

Despite these advancements, CVSS 4.0 presents certain challenges. The increased complexity of the scoring system can lead to difficulties in implementation and interpretation. Our experiments revealed that the detailed metrics while providing more granularity, can also be overwhelming and require extensive domain knowledge to apply effectively. Additionally, the complexity of the CVSS 4.0 framework might hinder its adoption among practitioners who are accustomed to the more straightforward CVSS 3.1 model[5]. Future research should explore methods to streamline the application of CVSS 4.0, possibly through automated tools or simplified guidelines that can facilitate its use in practical scenarios.

5.2. Deep Learning Prospects in Vulnerability Assessment

(1) Prospects:

The integration of deep learning models into the CVSS 4.0 framework shows considerable promise for enhancing vulnerability assessment. Our results indicate that deep learning techniques can automate vulnerability scoring, provide real-time updates, and achieve high accuracy in risk assessments. For example, models such as Support Vector Machines (SVM) and Gradient Boosting Decision Trees (GBDT) demonstrated strong performance in predicting vulnerability trends and identifying potential security threats.

(2) Unexpected Findings:

One unexpected finding was that while deep learning models generally improved scoring accuracy, they sometimes struggled with highly imbalanced datasets. Techniques like SMOTE and class weight adjustments helped to some extent, but they did not fully resolve the issue. This highlights a need for more advanced methods to handle data imbalance, which could be a crucial area for future research.

DOI: 10.54254/2755-2721/87/20241621

(3) Challenges:
Deep learning models also faced challenges such as data quality and the interpretability of results. Although deep learning algorithms showed high-performance metrics, the quality of the training data significantly affected the outcomes. Future research should focus on improving data collection methods and developing techniques to enhance the interpretability of deep learning models, making them more transparent and actionable for practitioners.

(4) Future Research Directions:
To address the identified challenges and build on the current findings, future research should focus on the following areas:

① Enhancing Data Quality: Developing methods to gather more comprehensive and high-quality vulnerability data.

② Improving Algorithm Robustness: Exploring advanced algorithms and techniques to better handle data imbalance and enhance model robustness.

③ Increasing Interpretability: Creating methods to improve the interpretability of deep learning models, making their decisions more transparent and easier to understand.

Optimizing CVSS 4.0 Implementation: Investigating ways to simplify the application of CVSS 4.0 metrics and streamline its use for both experts and practitioners.

6. Conclusion

This paper presents a comprehensive review of the evolution of the Common Vulnerability Scoring System (CVSS) and explores the integration of deep learning techniques into vulnerability assessment through CVSS 4.0. By detailing the advancements introduced in CVSS 4.0 and evaluating the potential of deep learning models, this study provides a robust framework for enhancing vulnerability assessment processes.

6.1. Practical Implications of Findings

6.1.1. Enhanced Vulnerability Assessment:

The adoption of CVSS 4.0, with its advanced metrics and nuanced scoring capabilities, offers a more precise and comprehensive approach to evaluating vulnerabilities. The introduction of the Threat metric group and the Scope metric enables security professionals to assess vulnerabilities with greater detail, addressing limitations of previous versions and providing a more accurate reflection of the risks faced by organizations. This improvement facilitates better prioritization of security measures and resource allocation, directly benefiting organizations' security postures.

6.1.2. Integration with Deep Learning Technologies:

The integration of deep learning models with CVSS 4.0 represents a significant advancement in automating and refining vulnerability assessments. Deep learning techniques, such as Support Vector Machines (SVM) and Gradient Boosting Decision Trees (GBDT), have demonstrated the ability to handle large datasets, identify patterns, and predict future vulnerabilities with high accuracy. These technologies can automate routine assessment tasks, provide real-time updates, and enhance the efficiency of security operations. This integration supports more proactive and effective security measures, helping organizations stay ahead of evolving threats.

6.2. Impact on the Field of Information Security

6.2.1. Proactive Security Measures

The combined use of CVSS 4.0 and deep learning techniques represents a shift towards more proactive security measures. By improving the accuracy of vulnerability assessments and enabling real-time updates, these methods allow organizations to anticipate and address potential threats before they can

DOI: 10.54254/2755-2721/87/20241621

be exploited. This proactive approach not only enhances immediate security defenses but also contributes to long-term risk management strategies.

6.2.2. Scalability and Efficiency

The proposed methods offer scalable solutions for vulnerability assessment across various organizational sizes and types. Automated scoring and analysis provided by deep learning models reduce the reliance on manual processes, thereby increasing the efficiency of vulnerability management. This scalability ensures that organizations of all sizes can benefit from advanced vulnerability assessment techniques, promoting broader adoption of best practices in information security.

6.2.3. Future Research Directions

The study identifies several areas for future research that can further enhance the effectiveness of CVSS 4.0 and deep learning applications in vulnerability assessment. These include:

① Enhancing Data Quality: Developing new methods for collecting and refining vulnerability data to ensure it is comprehensive and accurate.

② Improving Algorithm Robustness: Exploring advanced algorithms and techniques to address data imbalance issues and improve the robustness of deep learning models.

③ Increasing Interpretability: Creating approaches to make deep learning models' predictions more transparent and understandable for practitioners.

④ Optimizing CVSS 4.0 Implementation: Investigating ways to simplify and streamline the application of CVSS 4.0 metrics for practical use in diverse environments.

References

[1] Scarfone K, Mell P. An analysis of CVSS version 2 vulnerability scoring[C]//2009 3rd International Symposium on Empirical Software Engineering and Measurement. IEEE, 2009: 516-525.

[2] Gallon L, Bascou J J. Using CVSS in attack graphs[C]//2011 Sixth International Conference on Availability, Reliability, and Security. IEEE, 2011: 59-66.

[3] Ruohonen J. A look at the time delays in CVSS vulnerability scoring[J]. Applied Computing and Informatics, 2019, 15(2): 129-135.

[4] Costa J C, Roxo T, Sequeiros J B F, et al. Predicting CVSS metric via description interpretation[J]. IEEE Access, 2022, 10: 59125-59134.

[5] Gallon L. On the impact of environmental metrics on CVSS scores[C]//2010 IEEE Second international conference on social computing. IEEE, 2010: 987-992.

Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241618

The analysis of the impact of different supply chain factors using statistical perspective

Tianqin Xiong

Sichuan University - Pittsburgh Institute, Sichuan University, Chengdu, 610207, China

2021141520136@stu.scu.edu.cn

Abstract. Supply Chain Management (SCM) is an essential part of modern business and involves the coordination of procurement, production, and logistics. Effective SCM allows organizations to remain competitive and reduce costs. With the growth of the global marketplace, SCM is becoming increasingly important for businesses to compete. The most important aspect of SCM is to understand the components of the supply chain, which form a network of organizations involved in value-added processes and activities. The supply chain network consists of multiple components, indicating that the success of an organization's SCM depends on the combination of multiple influencing factors. This paper utilizes the supply chain data of several well-known companies, uses a dozen variables such as inventory turnover ratio, lead time, and supplier count as independent variables, and analyzes the effects of these different supply chain influencing factors on operational efficiency, supply chain resilience, and supplier relationship through multiple linear regression models. This study finds that many different supply chain factors, like transportation cost efficiency, delivery time, and others, have the greatest impact on the above-mentioned factors. This analysis may help companies to comprehensively assess their supply chain maturity and strategically enhance their supply chain attributes to achieve their business development goals.

Keywords: Supply Chain Management, Multiple Linear Regression, Operational Efficiency, Supply Resilience.

1. Introduction

Supply Chain Management (SCM) is a key component of modern business strategy, encompassing the coordination and integration of multiple activities such as procurement, production, and logistics. For an organization, efficient supply chain management can maintain competitiveness, reduce costs, and ensure timely delivery of products to consumers [1]. The complete supply chain is a network of organizations that are linked upstream and downstream and are involved in different processes and activities that generate value in the form of products and services provided to the final consumer. And there are many types of methods to perform supply chain management and enhancement, which include Agile SCM, Lean Manufacturing, and Cross-Docking, etc. These supply chain management methods can positively affect inventory levels, improve order fulfillment, and increase customer satisfaction by enhancing the company's benefits. Many elements may influence the supply chain, and various attributes of the supply chain help assess its performance [2]. Understanding the specific impact of different SCM

DOI: 10.54254/2755-2721/87/20241618

practices on operational metrics and relational aspects is critical for organizations to achieve sustainable growth and increased resilience in a dynamic business environment [3].

The purpose of this study is to explore how different supply chain factors affect a firm's operational efficiency, supply chain resilience, and supplier relationships. This paper will find out which practices of supply chain management have a significant impact on operational efficiency, resilience, and supplier relationships. And it will elaborate on the main factors that affect operational efficiency, resilience, and supplier relationships. Operational efficiency represents the supply chain's ability to reduce waste, including time, cost, and resources, during the production and distribution process It is a crucial attribute to evaluate, as a highly operationally efficient supply chain can bring great benefits to an enterprise [4]. Supply chain resilience, on the other hand, is the measure of the supply chain system's capability to respond to the risk factors that can threaten the supply chain and its ability to bring the supply chain back to its previous condition or even improve it in light of various levels of risk [5-6]. This attribute is the agility of the supply chain to counter a disruption, that is, to replenish supply, and this has a direct effect on the robustness of a company's supply chain [7]. Supplier relationships focus on the level of connections between the enterprise and its suppliers. Thus, it is vital for an enterprise to build a strong partnership with its suppliers in order to gain a competitive advantage. Beneficial supplier relationships lead to better product quality, lower costs, and faster delivery, which in turn leads to innovation, increased competitiveness, and business opportunities for companies. The purpose of this study is to reveal the precise functions of these approaches in increasing firms' operational efficiency, enhancing supply chain resilience, and optimizing supplier relationships. The research aims to determine and evaluate the particular effects of supply chain factors on operational efficiency, to examine the impact of these factors on supply chain resilience in the response of the supply chain to disturbances and changes, and to investigate the impact of the supply chain factors in building and sustaining cooperation with suppliers. This research work does not only seek to contribute theoretical knowledge but also to help organizations in choosing and applying supply chain management strategies.

2. Methods

2.1. Data Source

In order to ensure the authority and accuracy of the data source, the supply chain event database of several well-known companies is used as the data source in this paper [8]. The database captures dozens of supply chain factors and three supply chain attribute scores from these companies. In order to ensure the comprehensiveness of the data and the accuracy of the modeling, 999 sets of supply chain data were collected from this database. Some supply chain factors with missing data are eliminated, resulting in a final dataset of 15 supply chain factors.

2.2. Variable Description

Table 1 show the independent and dependent variables in the 18 variables used in the study.

Table 1. Independent and Dependent variables of the model

Independent Variable	Dependent Variable
Inventory Turnover Ratio	Operational Efficiency Score
Lead Time (days)	Supply Chain Resilience Score
Supplier Count	Supplier Relationship Score
Order Fulfillment Rate (%)	-
Customer Satisfaction (%)	-
Supply Chain Agility: High (ref = Medium)	-
Supplier Lead Time Variability (days)	-
Inventory Accuracy (%)	-
Transportation Cost Efficiency (%)	-

DOI: 10.54254/2755-2721/87/20241618

Table 1. (continued).

Supply Chain Integration Level: High (ref = Medium)	-
Supply Chain Complexity Index: Low (ref = Medium)	-
Supply Chain Complexity Index: High (ref = Medium)	-
Cost of Goods Sold (COGS)	-
Revenue Growth Rate out of (15)	-
Supply Chain Risk (%)	-

As shown in Table 1, 18 variables are categorized into independent and dependent variables. Independent variables are the supply chain factors of these firms, and dependent variables are the attribute scores of the supply chain that can be assessed as influenced by these factors.

2.3. Mathematical Statistics Method

The mathematical statistical method used in this study is the Multiple Linear Regression Model.

The relationship between the response variables y_i and the predictor variables x_i can be modeled through a linear regression, where there are p predictor variables $x_1, x_2, \cdots, x_p$ and a single response variable y_i.

The relationship between the response variable y_i and the predictor variables x_i can be modeled with a multiple linear regression model. In this model, there are p predictor variables $x_1, x_2, \cdots, x_p$ and a single response variable y_i. The mathematical expression for the multiple linear regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad (1)$$

where ε_i is random error item, $\beta_0, \beta_1, \cdots, \beta_p$ are regression coefficient. And this equation represents a multiple linear regression model [9].

3. Results and Discussion

3.1. Descriptive Analysis

As shown in Table 2, several metrics exhibit significant characteristics., in which Supplier Count, shows great variability, ranging from 100 to 1,300,000, with a mean of 43,202.98 and a standard deviation of 213,665.67. This indicates that there are significant differences in the number of suppliers managed by different firms, with some firms relying on a small number of suppliers, while others manage a large network of suppliers.

Cost of Goods Sold (COGS) also shows significant variation ranging from 300 to 1500, with a mean of 719.02 and a standard deviation of 271.88. This demonstrates the variability in the cost of goods sold by different firms, potentially attributed to their business model and size. On the contrary, Order Fulfillment Rate and Inventory Accuracy show consistency. The Order Fulfillment Rate ranges from 87% to 99%, with an average of 91.73% and a standard deviation of 2.89, indicating that most companies are able to achieve a high order fulfillment rate. Inventory accuracy ranged from 95% to 99%, with an average of 97.41% and a standard deviation of only 1.17, indicating that most companies excel in inventory management.

Table 2. Descriptive Statistical Analysis Results

Variable	N	Minimum Value	Maximum Value	Average Value	Standard Deviation
Inventory Turnover Ratio	999	1	50	6.33	5.48
Lead Time (days)	999	2	22	11.45	4.20
Supplier Count	999	100	1300000	43202.98	213665.671
Order Fulfillment Rate (%)	999	87	99	91.73	2.89

DOI: 10.54254/2755-2721/87/20241618

Table 2. (continued).

Customer Satisfaction (%)	999	85	94	89.21	2.27
Supplier Lead Time Variability (days)	999	1	10	3.28	1.59
Inventory Accuracy (%)	999	95	99	97.41	1.17
Transportation Cost Efficiency (%)	999	80	92	87.15	2.38
Cost of Goods Sold (COGS)	999	300	1500	719.02	271.88
Revenue Growth Rate out of (15)	999	8	20	10.83	1.95
Supply Chain Risk (%)	999	3	15	8.95	2.86
Operational Efficiency Score	999	75	90	83.45	2.31
Supply Chain Resilience Score	999	80	95	88.09	2.86
Supplier Relationship Score	999	78	90	83.88	2.85

Table 3 is a correlation analysis. Several key metrics with significant characteristics have been found:

The correlation coefficient between Order Fulfillment Rate and Supply Chain Resilience Score is as high as 0.860, with a significance level of $p < 0.001$, which suggests that a high Order Fulfillment Rate is usually accompanied by a high Supply Chain Resilience Score, implying that supply chains are more adaptive and flexible in the presence of efficient order processing. This suggests that when order processing efficiency is high, a high Supply Chain Resilience Score usually follows, indicating that the supply chain is more adaptable and flexible .

The correlation coefficient between Supply Chain Risk and Supply Chain Resilience Score is -0.873 with a significance level of $p < 0.001$, which shows a strong negative correlation between the two, which indicates the importance of reducing supply chain risk in order to improve supply chain resilience.

There is also a strong positive correlation between Transportation Cost Efficiency and Supply Chain Resilience Score, with a correlation coefficient of 0.809 and a significance level of $p < 0.001$, indicating that high Transportation Cost Efficiency is usually accompanied by high Supply Chain Resilience Score, suggesting that optimizing transportation costs can help improve supply chain resilience. Score, indicating that optimizing transportation costs helps to improve supply chain resilience and responsiveness.

Finally, there is a positive correlation between Customer Satisfaction and Order Fulfillment Rate, which indicates that an increase in Order Fulfillment Rate usually leads to an increase in Customer Satisfaction, which plays an important role in improving the competitiveness and market position of enterprises.

Table 3. Correlation Analysis Results

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1 Inventory Turnover Ratio	1													
2 Lead Time (days)	-0.013	1												
3 Supplier Count	-0.012	-0.037	1											
4 Order Fulfillment Rate (%)	0.032	-0.783***	0.031	1										
5 Customer Satisfaction (%)	0.072*	-0.095**	0.017	0.104**	1									
6 Supplier Lead Time Variability (days)	-0.008	0.804***	-0.030	-0.696***	-0.071*	1								
7 Inventory Accuracy (%)	-0.006	-0.672***	-0.027	0.499***	0.073*	-0.726***	1							
8 Transportation Cost Efficiency (%)	0.010	-0.647***	-0.024	0.622***	0.069*	-0.788***	0.625***	1						
9 Cost of Goods Sold (COGS)	0.023	-0.248***	0.036	0.601***	0.070*	-0.376***	0.284***	0.355***	1					
10 Revenue Growth Rate out of (15)	-0.003	-0.644***	0.118**	0.674***	0.055	-0.579***	0.492***	0.448***	0.535***	1				
11 Supply Chain Risk (%)	-0.016	0.853***	-0.026	-0.867***	-0.093**	0.795***	-0.557***	-0.712***	-0.429***	-0.621***	1			

DOI: 10.54254/2755-2721/87/20241618

Table 3. (continued).

12 Operational Efficiency Score	0.077*	-0.128***	0.017	0.153***	0.105**	-0.192***	0.184***	0.226***	0.139***	0.090**	-0.164***	1		
13 Supply Chain Resilience Score	0.051	-0.722***	-0.012	0.860***	0.097**	-0.756***	0.558***	0.809***	0.559***	0.498***	-0.873***	0.223***	1	
14 Supplier Relationship Score	0.029	-0.627***	-0.041	0.786***	0.077*	-0.646***	0.619***	0.613***	0.624***	0.508***	-0.757***	0.191***	0.839***	1

***p < 0.001; **p < 0.01; *p < 0.05.

3.2. Inferential Analysis

This paper selects 999 cases from the database, assuming a linear relationship between the three dependent variables and the 14 independent variables, and picks 14 groups of influences with low covariance for analysis. In this paper, those cases were used as training samples and analyzed by multiple linear regression using SPSS. This paper will analyze the three dependent variables: Operational Efficiency Score, Supply Chain Resilience Score and Supplier Relationship Score separately. And analyze the results of their regression, i.e., parameters such as regression coefficients, to derive the factors affecting these three parameters that can be assessed for the level of the supply chain. This will be of great significance for the company as it improves its supply chain in the future.

The multiple linear regression coefficient estimates were obtained through SPSS, and the analysis results for the three different supply chain attribute scores are shown in Tables 4, 5, and 6.

Table 4. Regression Analysis Results (Operational Efficiency Score)

	Unstandardized Coefficient		β	t	P	Covariance Statistics	
	B	SE				Tolerances	VIF
Constant	46.236	16.090		2.874	0.004		
Inventory Turnover Ratio	0.028	0.013	0.066	2.155	0.031	0.988	1.012
Lead Time (days)	0.068	0.048	0.124	1.430	0.153	0.126	7.983
Supplier Count	-0.004	0.033	-0.004	-0.120	0.905	0.972	1.029
Order Fulfillment Rate (%)	0.037	0.067	0.047	0.558	0.577	0.134	7.463
Customer Satisfaction (%)	0.086	0.032	0.084	2.699	0.007	0.972	1.029
Supply Chain Agility: High (ref = Medium)	-0.369	0.471	-0.058	-0.784	0.422	0.170	5.889
Supplier Lead Time Variability (days)	-0.088	0.116	-0.061	-0.758	0.448	0.146	6.842
Inventory Accuracy (%)	0.133	0.143	0.067	0.926	0.355	0.178	5.623
Transportation Cost Efficiency (%)	0.146	0.056	0.150	2.627	0.009	0.286	3.498
Supply Chain Integration Level: High (ref = Medium)	0.573	0.420	0.098	1.363	0.173	0.183	5.476
Supply Chain Complexity Index: Low (ref = Medium)	0.168	0.302	0.028	0.556	0.579	0.377	2.653
Supply Chain Complexity Index: High (ref = Medium)	-0.076	0.262	-0.014	-0.291	0.771	0.403	2.478
Cost of Goods Sold (COGS)	0.001	0.000	0.062	1.256	0.210	0.386	2.593
Revenue Growth Rate out of (15)	-0.070	0.057	-0.059	-1.235	0.217	0.408	2.452
Supply Chain Risk (%)	-0.017	0.073	-0.021	-0.229	0.819	0.114	8.757
R2				0.078			
R2 (After adjustment)				0.064			
F				5.571***			

***p < 0.001.

According to the data analysis results in Table 4, the three variables with the largest absolute values of standardized regression coefficients are Transportation Cost Efficiency (%) (0.150), Lead Time (days) (0.124), and Supply Chain Integration Level (0.098), and all three variables have positive coefficients.

DOI: 10.54254/2755-2721/87/20241618

The coefficients of all three variables are positive. This indicates that Transportation Cost Efficiency, Lead Time and Supply Chain Integration Level have a significant positive effect on Operational Efficiency Score. Therefore, these three supply chain factors are identified as the main factors affecting the Operational Efficiency Score. In order to improve the level of Operational Efficiency, companies should focus on improving Transportation Cost Efficiency, shortening Lead Time, and enhancing Supply Chain Integration Level.

Table 5. Regression Analysis Results (Supply Chain Resilience Score)

	Unstandardized Coefficient		*β*	*t*	*P*	Covariance Statistics	
	B	*SE*				Tolerances	VIF
Constant	27.051	5.668		4.773	0.000		
Inventory Turnover Ratio	0.012	0.005	0.024	2.698	0.007	0.988	1.012
Lead Time (days)	0.059	0.017	0.087	3.537	0.000	0.126	7.983
Supplier Count	-0.021	0.012	-0.016	-1.805	0.071	0.972	1.029
Order Fulfillment Rate (%)	0.441	0.024	0.446	18.718	0.000	0.134	7.463
Customer Satisfaction (%)	0.001	0.011	0.001	0.078	0.938	0.972	1.029
Supply Chain Agility: High (ref = Medium)	-0.337	0.166	-0.043	-2.035	0.042	0.170	5.889
Supplier Lead Time Variability (days)	-0.022	0.041	-0.012	-0.526	0.599	0.146	6.842
Inventory Accuracy (%)	-0.106	0.051	-0.044	-2.104	0.036	0.178	5.623
Transportation Cost Efficiency (%)	0.387	0.020	0.322	19.763	0.000	0.286	3.498
Supply Chain Integration Level: High (ref = Medium)	1.238	0.148	0.171	8.361	0.000	0.183	5.476
Supply Chain Complexity Index: Low (ref = Medium)	0.441	0.106	0.059	4.146	0.000	0.377	2.653
Supply Chain Complexity Index: High (ref = Medium)	0.298	0.092	0.059	4.311	0.000	0.403	2.478
Cost of Goods Sold (COGS)	0.002	0.00	0.151	10.756	0.000	0.386	2.593
Revenue Growth Rate out of (15)	-0.336	0.020	-0.229	-16.780	0.000	0.408	2.452
Supply Chain Risk (%)	-0.307	0.026	-0.308	11.915	0.000	0.114	8.757
R^2				0.925			
R^2 *(After adjustment)*				0.924			
F				810.580***			

***p < 0.001.

According to the results of the data analysis in Table 5, the three variables with the largest absolute values of standardized regression coefficients are Order Fulfillment Rate (%) (0.446), Transportation Cost Efficiency (%) (0.322), and Supply Chain Risk (0.308). The coefficients of these three variables are significantly higher than the coefficients of other factors, and among them the standardized regression coefficients of Order Fulfillment Rate and Transportation Cost Efficiency are positive, while the standardized regression coefficient of Supply Chain Risk is negative. This indicates that Order Fulfillment Rate and Transportation Cost Efficiency have a very significant positive effect on Supply Chain Resilience, while Supply Chain Risk has a very significant negative effect on Supply Chain Resilience. negative effect on Supply Chain Resilience. Therefore, these three supply chain factors were

DOI: 10.54254/2755-2721/87/20241618

identified as the main factors affecting the Supply Chain Resilience Score. In order to improve the Supply Chain Resilience, enterprises should focus on improving the Order Fulfillment Rate and Transportation Cost Efficiency in the supply chain, and effectively reducing Supply Chain Risk. By optimizing these key factors, enterprises can significantly improve the resilience and stability of their supply chains.

Table 6. Regression Analysis Results (Supplier Relationship Score)

	Unstandardized Coefficient		*β*	*t*	*P*	Covariance Statistics	
	B	*SE*				Tolerances	VIF
Constant	33.305	9.127		3.649	0.000		
Inventory Turnover Ratio	0.006	0.007	0.11	0.757	0.449	0.988	1.012
Lead Time (days)	0.063	0.027	0.093	2.340	0.020	0.126	7.983
Supplier Count	-0.039	0.019	-0.030	-2.082	0.038	0.972	1.029
Order Fulfillment Rate (%)	0.423	0.038	0.429	11.140	0.000	0.134	7.463
Customer Satisfaction (%)	-0.020	0.018	-0.016	-1.114	0.266	0.972	1.029
Supply Chain Agility: High (ref = Medium)	1.657	0.267	0.212	6.209	0.000	0.170	5.889
Supplier Lead Time Variability (days)	-0.109	0.066	-0.061	-1.657	0.098	0.146	6.842
Inventory Accuracy (%)	0.250	0.081	0.103	3.077	0.002	0.178	5.623
Transportation Cost Efficiency (%)	-0.086	0.032	-0.072	-2.718	0.007	0.286	3.498
Supply Chain Integration Level: High (ref = Medium)	-1.133	0.238	-0.157	-4.755	0.000	0.183	5.476
Supply Chain Complexity Index: Low (ref = Medium)	1.889	0.171	0.253	11.032	0.000	0.377	2.653
Supply Chain Complexity Index: High (ref = Medium)	0.184	0.149	0.027	1.236	0.217	0.403	2.478
Cost of Goods Sold (COGS)	0.003	0.000	0.304	13.388	0.000	0.386	2.593
Revenue Growth Rate out of (15)	-0.342	0.032	-0.234	-10.617	0.000	0.408	2.452
Supply Chain Risk (%)	-0.385	0.042	-0.386	-9.263	0.000	0.114	8.757
R^2				0.805			
R^2 *(After adjustment)*				0.802			
F				270.712***			

***p < 0.001.

According to the results of the data analysis in Table 6, the three variables with the largest absolute values of standardized regression coefficients are Order Fulfillment Rate (%) (0.429), Supply Chain Risk (%) (0.386) and Cost of Goods Sold (COGS) (0.304). The standardized regression coefficients of Order Fulfillment Rate and Cost of Goods Sold (COGS) are positive, while the standardized regression coefficient of Supply Chain Risk is negative. This indicates that Order Fulfillment Rate and Cost of Goods Sold (COGS) have a significant positive effect on Supplier Relationship while Supply Chain Risk has a significant negative effect on Supplier Relationship. Therefore, these three supply chain factors were identified as the main factors affecting Supplier Relationship Score. In order to improve the Supplier Relationship level of a company, companies should focus on improving Order Fulfillment Rate and Cost of Goods Sold (COGS) and reducing Supply Chain Risk.

DOI: 10.54254/2755-2721/87/20241618

4. Conclusion

This research aims to analyze the effects of various supply chain factors on operational efficiency, supply chain resilience, and supplier relationships through multiple linear regression on 999 cases of various companies' supply chains. The findings of this research reveal that the variables that are critical to supply chain management affect the performance of firms in the supply chain.

First, operational efficiency, transportation cost efficiency, delivery time, and supply chain integration level are the main factors in the supply chain process. The results indicate that the transportation cost efficiency, delivery time and supply chain integration level are all beneficial to the improvement of enterprises' operational efficiency. Second, the factors that affect supply chain resilience include order fulfillment rate, transportation cost efficiency, and supply chain risk. The findings of the study indicate that the order fulfillment rate and transportation cost efficiency positively and significantly influence supply chain resilience, while the supply chain risk negatively influences supply chain resilience. Finally, the supplier relationship is influenced by the order fulfillment rate, cost of the goods sold, and supply chain risk. The result of the analysis implies that enhancing the order fulfillment rate, managing the cost of the goods sold, and lowering the supply chain risk are the key factors for enhancing the supplier relations.

This study provides empirical evidence and specific suggestions on how enterprises can optimize supply chain management through statistical methods. Therefore, by increasing the transportation cost efficiency, delivery time, supply chain integration level, order fulfillment rate, cost of goods sold, and supply chain risk, companies can greatly increase their operational efficiency as well as supply chain resilience and supplier relationships. This makes it easier for enterprises to increase their competitiveness and achieve stable development in a competitive and constantly changing market environment. The results of this study are based on the current market and economic environment, and future changes in the market environment may affect the applicability of these findings. Therefore, it is suggested that future market trends and changes may further validate and expand the findings of this study.

References

[1] Hendricks, K. B., Singhal, V. R., & Stratman, J. K. (2007). The impact of enterprise systems on corporate performance: A study of ERP, SCM, and CRM system implementations. Journal of Operations Management, 25(1), 65–82. https://doi.org/10.1016/j.jom.2006.02.002

[2] George, J., & Pillai, V. M. (2019). A study of factors affecting supply chain performance. Journal of Physics: Conference Series, 1355, 012018. IOP. https://doi.org/10.1088/1742-6596/1355/1/012018

[3] Abdul Halim, Z. (2016). The Moderating Effect of Supply Chain Role on the Relationship between Supply Chain Practices and Performances. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2882965

[4] Dharmapala, P. S. (2008). Adding value in supply chains by improving operational efficiency using data envelopment analysis: a case from published data. International Journal of Applied Management Science, 1(2), 160. https://doi.org/10.1504/ijams.2008.021099

[5] Abdel-Basset, M., Gunasekaran, M., Mohamed, M., & Chilamkurti, N. (2019). A Framework for Risk assessment, Management and evaluation: Economic Tool for Quantifying Risks in Supply Chain. Future Generation Computer Systems, 90(2), 489–502. https://doi.org/10.1016/j.future.2018.08.035

[6] Ponis, S. T., & Koronis, E. (2012). Supply Chain Resilience: Definition Of Concept And Its Formative Elements. Journal of Applied Business Research (JABR), 28(5), 921. https://doi.org/10.19030/jabr.v28i5.7234

[7] Bakshi, N., & Kleindorfer, P. (2009). Co-opetition and Investment for Supply-Chain Resilience. Production and Operations Management, 18(6), 583–603. https://doi.org/10.1111/j.1937-5956.2009.01031.x

[8] Supply Chain Management. (2024). https://www.kaggle.com/datasets/lastman0800/supply-

DOI: 10.54254/2755-2721/87/20241618

chain-management
[9] Marill, K. A. (2004). Advanced Statistics: Linear Regression, Part II: Multiple Linear Regression. Academic Emergency Medicine, 11(1), 94–102. https://doi.org/10.1111/j.1553-2712.2004.tb01379.x