Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

Application and Analysis in Machine Learning Algorithms for Financial Fraud Detection

Lebin Hu[1,*], Xinyang Liu[2], Changwen Luan[3]

[1] School of Business, Jiangnan University, Jiangsu, China

[2] College of Information and Electrical Engineering, China Agricultural University, Beijing, China

[3] Faculty of Engineering, The Chinese University of Hong Kong, Hong Kong, China

* Corresponding Author: 1089210228@stu.jiangnan.edu.cn

Abstract. This study provides a comprehensive examination of machine learning models and deep learning algorithms for detecting financial fraud. We begin by introducing the basic concepts of commonly used algorithms, discussing their strengths and limitations in identifying fraudulent activities. The focus is on the unique capabilities of deep learning models, particularly Graph Neural Networks (GNNs), in handling large-scale and high-dimensional data, and their proficiency in enhancing fraud detection accuracy and efficiency through automated feature extraction. Furthermore, the study evaluates and contrasts the efficacy of various models in real-world scenarios, highlighting their performance in improving recall, precision, and overall robustness. We anticipate future advancements in fraud detection technology, including dynamic graph modeling and cross-domain transfer learning. Through detailed analysis and empirical validation, this study not only enhances our understanding of existing technologies but also offers valuable insights and guidance for research and practice in the field of financial security. The empirical results demonstrate that deep learning models, especially the Residual-Layered-Camouflage Graph Neural Network (RLC-GNN) model, significantly outperform traditional methods in terms of accuracy, recall, area under the curve (AUC) and F1 scores, indicating a promising direction for future developments in financial fraud detection.

Keywords: Financial Fraud Detection; Machine Learning; Deep Learning; Graph Neural Networks (GNNs).

1. Introduction

The advent of technologies like big data, artificial intelligence (AI), and blockchain has revolutionized Internet financial services and enhanced efficiency but it has also introduced new fraud risks [1]. Traditional rule-based systems and machine learning models face challenges in detecting these frauds due to their inability to adapt quickly to new fraud patterns, handle non-linear complex relationships, and process high-dimensional data [2]. Moreover, they lack scalability and efficiency in processing big data for real-time fraud detection. To address these issues, there is a growing interest in deep learning and neural networks. These advanced techniques can automatically adapt, process large-scale data, and excel in pattern recognition as well as feature recognition. They also offer end-to-end learning from raw data, improving efficiency and accuracy in large-scale data processing, making them increasingly valuable in financial fraud detection.

Financial fraud detection is a continuously evolving field. From the early evaluation methods that relied on expert systems to today's deep learning technology, the development of related technologies has gone through several important stages [3]. The first stage is based on expert system evaluation. This type of method relies on prior knowledge. Although it can provide effective detection for specific fraudulent behaviors, it is costly and inefficient. With the increase in data volume and computing power, traditional machine learning algorithms have entered this field and become the mainstream method in the second stage. These algorithms improve the efficiency and generalization of detection through automated processing, but they still have certain limitations in processing complex data structures. Entering the third stage, the effect of fraud identification has been notably enhanced. Deep

learning can not only automate feature extraction but also has stronger adaptive capabilities and can identify more complex and hidden fraud patterns [3]. In existing research, methods such as autoencoders, graph embedding algorithms for node neighborhood exploration via Random Walks (Node2Vec), and Graph Neural Networks (GNNs) are widely used in the detection of financial frauds. Specific application cases include the successful application of feedback mechanism hybrid models and convolution neural networks (CNNs) in financial fraud identification [4]. These technologies not only improve the accuracy of detection but also show strong capabilities in processing large-scale data and complex network structures. The conclusion indicates that the fraud detection system utilizing deep learning technology has excellent processing capabilities for high-dimensional data, significantly enhancing the accuracy and efficiency of detection. The practice of these advanced methods in the financial field not only improves the level of risk management but also provides a solid foundation for further study and innovation.

The primary objective of this study is to conduct an extensive review of machine learning and deep learning algorithms for the detection of financial fraud. Initially, the study introduces the basic concepts of commonly used algorithms for fraud identification. Aiming to identify models with high classification accuracy, we analyze the principles of these algorithms and compare their advantages and disadvantages, with a strong emphasis on the strengths of deep learning models in fraud detection, which are pivotal in the age of big data. This study aims to summarize mainstream fraud detection algorithms and related works, discussing optimization methods for detection models to guide and inform researchers in the field. The paper is structured as follows: First, we introduce basic machine learning algorithms, such as Decision Trees and Support Vector Machines (SVM), along with their principles and the drawbacks of these models. In Section 2, we clarify the advantages of deep learning models, particularly those based on distributed big data approaches like GNNs, for fraud detection. Section 3 presents relevant indicators for evaluating model accuracy and compares the performance of these models in different contexts, demonstrating the specific types of fraud each is proficient at identifying. Finally, Section 4 concludes with relevant findings and synthesizes the future prospects of fraud detection methods.

2. Method

2.1. Dataset Description and Preprocessing

The research in financial fraud forecasting relies predominantly on datasets from financial institutions and internet companies. The first data set is a publicly available credit card fraud data set. It contains 1000000 samples with 8 key features. The fraudulent transactions constitute 2.96% of the total data set [3]. The second data set under scrutiny in this study originates from a Chinese online financial services provider, consisting of 192,586 entries, of which 4,375 are identified as fraudulent cases [4]. This data set consists of more than 60 variables, including financial status and income levels. The third data set is a fusion of two datasets: the Yelp Public data set and an open-source data set from Amazon. The Yelp data set includes 45,954 entries with 6,663 instances of fraud, while the Amazon data set comprises 11,944 entries with 1,134 instances of fraudulent activity [5].

2.2. Proposed Approach

This study aims to explore existing fraud detection models through a comprehensive literature review of papers employing machine learning and deep learning methods for fraud identification. The research begins with an introduction to fundamental models, specifically SVM and Decision Tree algorithms. Their loss function expressions and the general steps for parameter estimation are presented. Next, we delve into deep learning algorithms such as CNNs and GNNs, highlighting the characteristics of graph-embedding algorithms and the construction of GNNs. In the third section, we analyze the traits of various fraud detection models. We examine a hybrid classification model based on K-means and Decision Trees, as well as a classification model utilizing graph embedding algorithms for feature selection, summarizing their innovative points and comparing metrics such as

precision and recall scores. Additionally, we discuss the Residual-Layered-Camouflage Graph Neural Network (RLC-GNN) model for the identification of fraud, describing the problems it addresses and analyzing its convergence rate and recall scores. Finally, in the fourth section, we synthesize the analysis results of fraud detection algorithms, comparing their advantages and disadvantages and considering their applicability. The future development directions of fraud detection algorithms are also introduced. This study provides a detailed overview of current methodologies, offering valuable insights and guidance for researchers in the field. Fig.1 illustrates the research methodology.

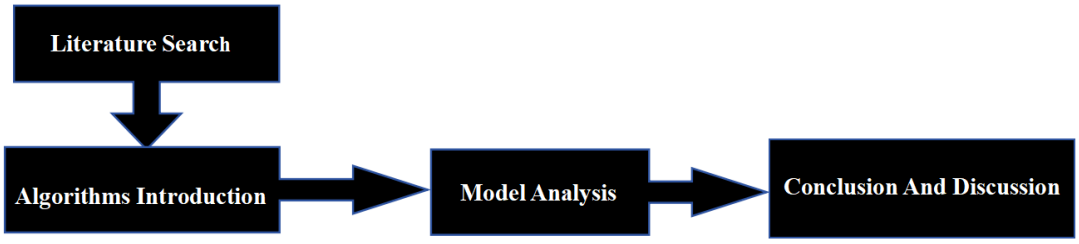

Figure 1. The Methodology of The Research

2.2.1. Introduction to Machine Learning and Deep Learning Algorithms

SVM represents a machine learning-based approach aimed at discovering the most effective dividing hyperplane that aligns with the necessary criteria for classification. The general expression of SVM can be expressed as: For a classification problem with a training set, the model needs to find a hyperplane, which can be expressed as:

$$f(x) = w^T x + b = 0 \tag{1}$$

where w $\in R^n$ and b $\in$ R. Specifically, SVM aims to find a decision boundary that not only correctly classifies the data into two categories but also maximizes the distance between this boundary and the nearest data points. This distance is referred to as the margin, and maximizing the margin enhances the model's generalization capability. Additionally, to measure the empirical risk, the soft margin loss function is used. SVM can also be applied to non-linearly separable data by employing the kernel trick, which maps the original data into a higher-dimensional space where a linear separating hyperplane can be found. Common kernel functions include the linear kernel and Radial Basis Function (RBF) kernel. In addition, a decision tree is a diagrammatic representation where the leaves signify different classes, while the internal nodes correspond to the evaluation of features [6,7]. To construct a decision tree, algorithms like C4.5 are employed. The C5.0 algorithm, an advancement of C4.5 by Quinlan, is founded on the decision tree principle. It offers enhanced efficiency and memory usage compared to C4.5. In the C5.0 algorithm, the partitioning of data samples is determined by the attribute that yields the greatest increase in information, known as the highest information gain field [8]. Information gain can be obtained through steps as follows:

$$I(S) = -\sum_{i=1}^{n} p(s_i) \log_2 p(s_i) \tag{2}$$

$$G(S,N) = I(S) - \sum \frac{I(S_n)}{S} I(S_n) \tag{3}$$

where I(S) is the entropy of the dataset and $\sum \frac{I(S_n)}{S} I(S_n)$ is conditional entropy for the data set given the variable N. C5.0 is capable of producing classification models in the form of decision trees; It also retains the capability to formulate classifiers in an easily comprehensible set of rules.

GNNs represent a category of neural networks specifically tailored for the analysis and learning of data structured in the form of graphs. They have attracted considerable interest because of their proficiency in discerning intricate patterns and connections within graphs, a common feature across diverse fields like social media networks and citation systems. GNNs operate by extending the principles of CNNs from regular grid data, like images, to irregular graph data. They achieve this by aggregating information from a node's neighbors and updating the node's representation through a series of learnable transformations. This process can be applied to capture multi-hop neighborhood information and produce a rich representation of the graph's structure and content. GNNs have been utilized across a broad spectrum of applications, encompassing tasks such as node categorization, link forecasting and graph synthesis. They have also been integrated into various applications across fields like computer vision, natural language processing. The advantage of GNNs lies in their ability to process graph-structured data which is highly irregular, making them a powerful tool for a multitude of tasks.

2.2.2. An Integrated Hybrid Model Using K-means and Decision Trees

This model utilizes a comprehensive hybrid methodology that leverages the capabilities of Apache Spark. The hybrid approach employs the K-Means clustering algorithm with the decision tree algorithm and is evaluated within the frameworks of Hadoop and Spark ecosystems. K-Means facilitates the division of the data into K coherent clusters, each centered around the nearest points. Subsequently, the decision tree algorithm is deployed on each cluster to develop a specific tree that sorts the instances into their respective categories, distinguishing between normal and anomalous patterns. The preference for K-Means is attributed to its computational efficiency. By pinpointing the most proximate clusters, K-Means paves the way for the decision tree to apply its classification rules, thereby accurately categorizing each data instance in the set as either typical or fraudulent. The detection process is depicted in Fig.2.

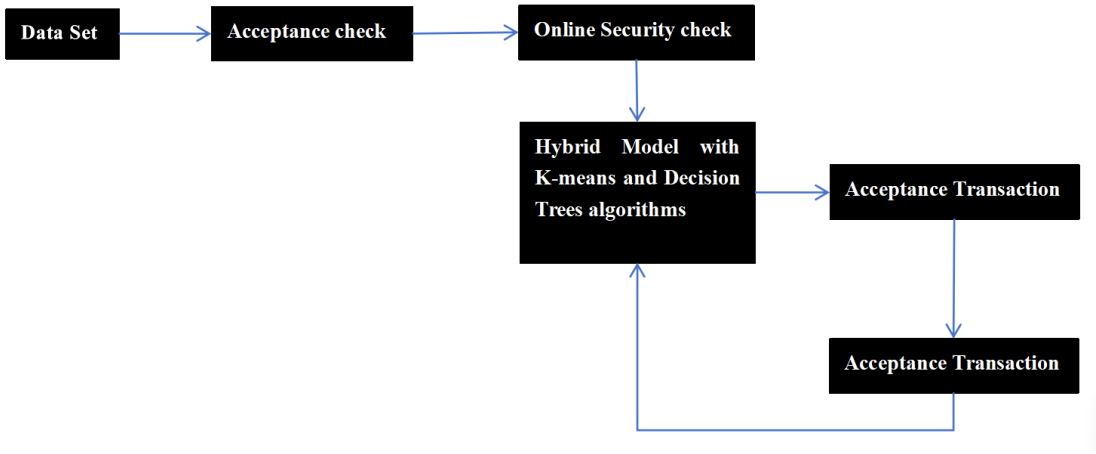

Figure 2. The Detection Process with The Hybrid Model

2.2.3. Node2Vec

Node2Vec is a network embedding algorithm that targets representing nodes in a graph as low-dimensional vectors so that these representations can be used for various tasks such as node classification, node clustering and link prediction. Node2Vec generates neighbor sequences of nodes by performing random walks on the graph and uses the Skip-Gram model to optimize the vector representations of these nodes. Its core advantage lies in the introduction of a flexible random walk strategy that combines Breadth First Search (BFS) and Depth First Search (DFS) to better capture the local and global structures of the network.

Specifically, Node2Vec introduces two important parameters to control the strategy of random walks: the first is the return parameter, which dictates the likelihood of revisiting the preceding node during the traversal. A higher p-value encourages the walk to return, tending to BFS, thereby capturing local structure; the second is the in-out parameter, which controls the probability of going to a node far

away from the previous node during the walk. A lower q value encourages the walk to move away from the previous node, tending to DFS, thereby capturing global structure. The application of Node2Vec mainly represents nodes in complex networks as low-dimensional vectors, so that these representations can be efficiently used by machine learning models to detect abnormal behavior. The application of Node2Vec in fraud identification algorithms such as SVM and decision trees enables feature selection much more accurate and efficient, thus boosting the model's performance. The detection process is depicted in Fig.3.

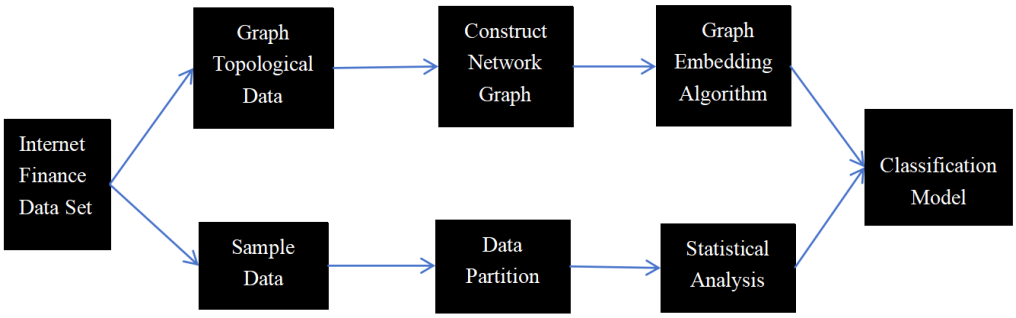

Figure 3. The Detection Process with Node2Vec

2.2.4. Residual Layered Camouflage-Resilient GNN (RLC-GNN)

RLC-GNN is a cutting-edge graph neural network technology designed for processing heterogeneous graph data, which is an improved version of Camouflage-Resilient GNN (CARE-GNN) [9]. Its uniqueness lies in its ability to accurately capture and utilize the complex and diverse node and edge relationships in the graph. Heterogeneous graphs are commonly found in real-world scenarios. Due to their complex structure and rich information, traditional graph neural networks are difficult to deal with effectively. RLC-GNN introduces the concept of meta-paths to define specific types of relationship links between nodes, thereby achieving a deep understanding of the complex structure and semantic information of heterogeneous graphs.

In RLC-GNN, meta-paths are the core mechanism that not only helps the model identify the associations between different types of nodes but also guides the extraction of contextual features to ensure that the model fully considers the semantic relationships of nodes during feature learning. Through a multi-perspective learning strategy, RLC-GNN can independently extract features from different meta-paths, and then automatically learn the optimal fusion solution through weighted aggregation to generate a comprehensive node representation. The application of RLC-GNN in fraud identification mainly uses its context awareness and ability to process heterogeneous graph data to identify and predict fraud in complex networks.

3. Result and Discussion

This section is dedicated to assessing the efficacy of the RLC-GNN model in the context of fraud detection, juxtaposing it against a variety of fraud detection algorithms rooted in machine learning and deep learning methodologies. We also explore the prospective trends in the field of financial fraud detection. In the initial section, we delve into the impact of different layers and parameter configurations on the loss function and convergence behavior of the RLC-GNN model. We further train and test the model using two distinct datasets, assessing the model's recall rates and analyzing its classification accuracy. The subsequent section of the study provides a comparative analysis and summary of the characteristics of the models in question, highlighting the current challenges and setting clear objectives for future research endeavors.

3.1. Experimental Results of RLC-GNN

The RLC-GNN significantly advances over the traditional CARE-GNN by implementing a multi-layer architecture complemented by a residual structure, specifically designed to enhance fraud detection performance. The application of RLC-GNN on Yelp and Amazon datasets yields substantial improvements, which we detail in Table 1. Specifically, for the Yelp data set, the model achieves a Recall of 76.68%, an area under the curve (AUC) of 85.44%, and an F1-Score of 70.03%, indicating good performance on this data set but suggesting potential misclassification issues with positive samples. In contrast, the Amazon data set demonstrates even more impressive results with a Recall of 91.83%, an AUC of 97.48%, and an F1-Score of 89.18%, indicating the model's high classification accuracy and well-balanced handling of positive and negative samples.

Table 1. The results in the data sets

Data set	Recall	AUC	F1-Score
Yelp Data set	76.68%	85.44%	70.03%
Amazon Data set	91.83%	97.48%	89.18%

The enhancements observed are largely due to the architectural design of RLC-GNN, which systematically rectifies inaccuracies through its multi-layer framework—a capability not available in single-layer architectures. This design integrates layered and residual configurations that facilitate a comprehensive and deep learning approach, optimizing the use of available neighboring data more effectively.

3.2. Discussion

As an advanced graph neural network model, RLC-GNN has shown significant advantages and limitations in the field of fraud detection. Its core advantage is that it can efficiently capture the complex relationships between nodes and edges in the graph structure. It is particularly good at processing data involving context and time series information. It elevates the precision and resilience of the model while also bolstering its capacity to discern dynamic behavioral patterns, thereby demonstrating strong adaptability and broad applicability across various graph datasets and fraud-related contexts. However, RLC-GNN also faces challenges such as high computational complexity, strong dependence on high-quality data, and poor model interpretability, which are also some of the problems faced by classic deep learning models. In summary, RLC-GNN is a fraud detection tool that combines high accuracy and versatility, but its application needs to carefully consider the requirements of computing resources and data quality.

In fraud identification, RLC-GNN has shown its significant advantages, such as strong context perception and excellent dynamic behavior modeling. It uses graph convolution technology to accurately grasp complex relationships, shows high generalization ability, significantly reduces false positive and false negative rates, and also solves the problem of feature disguise and relationship disguise to a certain extent [10]. For the problems of RLC-GNN in fraud identification, the following methods can be considered, such as using distributed and graph sampling technology to reduce computational complexity, data enhancement and self-supervised learning to meet data needs, hierarchical graph processing and efficient algorithms to improve scalability. These strategies further optimize the performance of RLC-GNN, making it better applicable to the field of fraud detection.

The future development of RLC-GNN will focus on model optimization, dynamic graph modeling, enhancing interpretability, improving robustness and security, and cross-domain transfer learning. It is widely used in financial fraud detection, e-commerce platform security, social network monitoring, network security protection, medical and health supervision, and supply chain management [10]. Through in-depth research and practice, RLC-GNN will continue to expand its performance boundaries and become the core driving force for multi-domain intelligent solutions.

4. Conclusion

As technology advances, traditional rule-based systems and machine learning models in financial fraud detection face significant challenges. These include difficulty in adapting to new fraud patterns, limited ability to handle nonlinear complex relationships, and insufficient scalability and efficiency for real-time processing in a big data environment. Deep learning and neural network technologies have emerged as promising solutions to these challenges. They offer automatic adaptation and processing of large-scale data, excelling in pattern recognition and feature extraction, thereby improving the efficiency and accuracy of detection for financial fraud through end-to-end learning. This study reviews existing machine learning and deep learning algorithms, highlighting the strengths of the RLC-GNN model in financial fraud detection. It demonstrates their superior classification accuracy and potential. The experimental outcomes indicate that the RLC-GNN model substantially surpasses conventional models, achieving significant enhancements in recall, AUC, and F1 scores. The RLC-GNN model's multi-layer and residual structures enhance its use of neighboring data, boosting accuracy and robustness. However, deep learning models also face challenges, including high computational complexity, stringent data quality requirements, and limited model interpretability. Future development will focus on model optimization, dynamic graph modeling, robustness improvement, and cross-domain transfer learning. Continued research and practice in deep learning technology will drive the development of intelligent solutions across multiple domains, providing stronger support for financial fraud detection and other applications.

5. Authors Contribution

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

References

[1] U. Paschen, C. Pitt, and J. Kietzmann. Artificial intelligence: Building blocks and an innovation typology. Bus. Horizons, 63(2) (2020), 147–155.

[2] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, G. Bontempi. Credit card fraud detection: a realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8) (2017), 3784-3797.

[3] T. Santosh, D. Ramesh. Machine Learning Approach on Apache Spark for Credit Card Fraud Detectio. Ingénierie des Systèmes d'Information, 25 (2020), 101-106.

[4] H. Zhou, G. Sun, S. Fu, L. Wang, J. Hu and Y. Gao. Internet Financial Fraud Detection Based on a Distributed Big Data Approach With Node2vec. In IEEE Access, 9 (2021), 43378-43386.

[5] Y. Zeng, & J. Tang. RLC-GNN: An Improved Deep Architecture for Spatial-Based Graph Neural Network with Application to Fraud Detection. Applied Sciences, 11(12) (2021), 5656.

[6] S. Ding, H. Huang, J. Yu, & H. Zhao. Research on the hybrid models of granular computing and support vector machine. Artificial Intelligence Review, 43(3) (2015), 565–577.

[7] A.R. Panhalkar, & D.D. Doye. A novel approach to build accurate and diverse decision tree forest. Evolutionary Intelligence, 15 (2022), 439–453.

[8] T.F. Ghanem, W.S. Elkilani, H.M. Abdul-Kader. A hybrid approach for efficient anomaly detection using metaheuristic methods. Journal of Advanced Research, 6(4) (2015), 609-619.

[9] N. Bandinelli, M. Bianchini, F. Scarselli. Learning long-term dependencies using layered graph neural networks. In Proceedings of the 2010 International Joint Conference on Neural Networks, (2010), 1–8.

[10] D. Yingtong and L. Zhiwei and S. Li and D. Yutong and P. Hao and Y. Philip. Enhancing Graph Neural Network-based Fraud Detectors against Camouflaged Fraudsters. ACM International Conference on Information & Knowledge Management, (2020), 315-324.