Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 12, No. 4, 2024, 487-495.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

FinFD-GCN: Using Graph Convolutional Networks for Fraud Detection in Financial Data

Mohamad Mahdi Yadegar and Hossein Rahmani*

School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran.

Article Info

Article History:
Received 03 August 2024
Revised 23 November 2024
Accepted 21 December 2024

DOI:10.22044/jadm.2024.14869.2585

Keywords:
Fraud Detection, Machine Learning, Credit Card, Graph Convolutional Networks, Graph Representation, Node Classification.

**Corresponding author:*
h_rahmani@iust.ac.ir (H. Rahmani).

Abstract

In recent years, new technologies have brought new innovations into the financial and commercial world, giving fraudsters many ways to commit fraud and cost companies big time. We can build systems that detect fraudulent patterns and prevent future incidents using advanced technologies. Machine learning algorithms are being used more for fraud detection in financial data. But the common challenge is the imbalance of the dataset which hinders traditional machine learning methods. Finding the best approach towards these imbalance datasets is the problem many of the researchers are facing when trying to use machine learning methods. In this paper, we propose the method called FinFD-GCN that use Graph Convolutional Networks (GCNs) for fraud detection in credit card transaction datasets. FinFD-GCN represents transactions as graph in which each node represents a transaction and each edge represents similarity between transactions. By using this graph representation FinFD-GCN can capture complex relationships and anomalies that may have been overlooked by traditional methods or were even impossible to detect with conventional approaches, thus enhancing the accuracy and robustness of fraud detection in financial data. We use common evaluation metrics and confusion matrices to evaluate the proposed method. FinFD-GCN achieves significant improvements in recall and AUC compared to traditional methods such as logistic regression, support vector machines, and random forests, making it a robust solution for credit card fraud detection. By using the GCN model for fraud detection in this credit card dataset we outperformed base models 5% and 10%, with respect to F1 and AUC, respectively.

1. Introduction

In recent years, technological advancements have led to almost all financial operations being conducted online and through computer systems [1]. E-commerce has played a significant role in the growth of businesses worldwide, prompting many large companies to carry out their financial transactions online. This increase has created opportunities for malicious attackers to employ various fraudulent methods, resulting in substantial costs for these companies [2]. Financial fraud is considered illegal or unethical behavior that allows an individual or group of individuals to gain financial benefits through unethical means [3].

These frauds can be observed in various domains, including credit card fraud [4], insurance fraud, money laundering, healthcare fraud, and securities and commodities fraud [5]. Given this market's exponential growth, it is expected that crimes in this domain will emerge and evolve. For instance, the Australian Competition Council reported a 190% increase in these crimes between 2017 and 2018. Similarly, in 2019, the UK reported a threefold increase in such crimes, highlighting the

growing importance of analyzing, detecting, and preventing these crimes [6].

A review of the relatively recent history of this field shows that fraudsters have consistently been active, causing significant financial losses to individuals and companies. For example, Visa, the second-largest card network globally, reports a fraud rate of less than 0.1% across all transactions. This performance is achieved through a multi-layered security infrastructure and an AI-based fraud detection system, preventing $25 billion in annual fraud [7].

This study focuses on detecting credit card fraud in online transactions, a major challenge in the financial sector due to its complex, evolving patterns and significant financial impact.

In this paper, we aim to implement Graph Convolutional Network (GCN) [8] model and analyze different metrics and results to find a more suitable method for detecting anomalies in financial data. As we know, data imbalance makes it challenging for machine learning models to learn abnormal patterns effectively, leading to suboptimal performance in detecting anomalies under test conditions. Using a proper model that can handle these kinds of datasets is an important task.

The proposed FinFD-GCN framework represents a significant advancement in fraud detection methodologies. The core innovation of our FinFD-GCN method lies in:

Novel Graph-Based Representation of Transactions: FinFD-GCN introduces an innovative approach by representing financial transactions as a graph structure. Each transaction is modelled as a node, and the edges capture the similarity between transactions based on predefined metrics. This graph-based representation enables the model to identify complex relational patterns that traditional methods often overlook, enhancing its ability to detect sophisticated fraud schemes.

The other unique elements outlined below demonstrate the importance and novelty of this work:

Superior Performance Metrics: Through extensive experimentation, FinFD-GCN demonstrates superior performance compared to traditional machine learning methods such as Logistic Regression, Random Forest, and Support Vector Machines. The model achieves higher recall and Area Under the Curve (AUC) scores, indicating its effectiveness in minimizing false negatives while maintaining strong discriminatory power between fraudulent and legitimate transactions.

Effective Handling of Imbalanced Datasets: Imbalanced datasets pose a significant challenge in fraud detection. FinFD-GCN addresses this issue by leveraging graph-based learning techniques that naturally incorporate minority class patterns through node connections and neighborhood aggregation. This capability ensures better detection of rare fraudulent transactions without extensive preprocessing or oversampling techniques.

Scalability and Robustness: FinFD-GCN is designed to handle large-scale financial transaction datasets efficiently. The framework's ability to process high-dimensional graph data and its use of batch normalization and dropout layers make it both scalable and robust. This ensures reliable performance even in real-world scenarios with millions of transactions and dynamic data environments.

In this paper, we compare these different models by applying the data to our GCN model and other classifiers to determine the extent to which our model results can be better. In Section 2, we review related work and examine papers in this field. In Section 3, we describe our model, the training process, and data generation. In Section 4, we compare the results obtained from the models and observe the impact of different sampling methods. Finally, in Section 5, we conclude and discuss future work.

2. Background and related works

Fraud detection has traditionally relied on classical machine learning algorithms, including Logistic Regression (LR), Random Forest (RF), and Support Vector Machines (SVM). These methods were instrumental in identifying fraudulent activities during the initial phases of automated fraud detection systems due to their simplicity and interpretability.

i. **Logistic Regression (LR):** LR has been widely used for its ease of implementation and efficiency in binary classification tasks. However, it assumes a linear relationship between features and the outcome, which often limits its ability to model complex fraud patterns. Research by Mahajan et al. [9] highlights its application in credit card fraud detection, but also points out its suboptimal performance when the data is highly imbalanced.

ii. **Random Forest (RF):** RF, a versatile ensemble method, improves classification accuracy by aggregating results from multiple decision trees. Its ability to

handle non-linearity and feature importance ranking has made it a popular choice. Nevertheless, RF can struggle with highly imbalanced datasets, leading to a bias toward the majority class, as emphasized by Yiu [10].

iii. **Support Vector Machines (SVM):** SVM has been employed for its robustness in handling high-dimensional data and complex decision boundaries. However, its reliance on careful parameter tuning and computational intensity in large datasets often limits scalability. Gyamfi and Abdulai [11] have discussed these challenges in the context of bank fraud detection.

Limitations of Traditional Methods

Despite their contributions, traditional methods exhibit notable limitations:

- **Inability to Model Relational Patterns:** LR, RF, and SVM analyze data in an independent feature space, making it difficult to capture the relational or networked structures often present in fraudulent transactions. This shortcoming hinders their ability to detect sophisticated fraud schemes that rely on these relationships.
- **Struggles with Imbalanced Data:** Fraud detection datasets typically exhibit significant class imbalances, with legitimate transactions vastly outnumbering fraudulent ones. Traditional methods are prone to overfitting to the majority class, resulting in higher false-negative rates and compromised detection accuracy for the minority (fraudulent) class [12].

Transition to Advanced Methods

The rise of graph-based approaches, such as Graph Convolutional Networks (GCNs), addresses these limitations by leveraging graph structures to represent relational patterns and incorporating techniques like graph sampling to mitigate imbalance issues. GCNs and other Graph Neural Networks (GNNs) have shown superior performance in fraud detection by capturing complex dependencies and temporal dynamics within transaction data, as highlighted in recent studies [13, 14].

The field of Graph Neural Networks (GNNs) has seen a surge in interest recently [13, 15]. Researchers have extended traditional neural network models like Convolutional Neural Networks (CNNs), which are designed for regular grid structures, to operate on irregular graph structures [14, 16]. Kipf and Welling's seminal work introduced Graph Convolutional Networks (GCN), a simplified version of GNNs, which achieved leading results on various benchmark graph datasets [8]. The ability of GCNs to manage structured data effectively makes them an excellent choice for identifying anomalies in financial datasets.

Ma et al. [17] provides an extensive review of contemporary deep learning techniques for graph anomaly detection, emphasizing the unique capabilities of GCNs in handling graph-structured data. It categorizes existing work based on the type of anomalies detected (node, edge, sub-graph, or graph level) and highlights the strengths and limitations of various approaches. The survey underscores the importance of leveraging graph representation learning for effective anomaly detection, particularly in complex datasets like those in financial transactions.

Dou et al. [18] addresses the challenges of detecting sophisticated fraudsters who use camouflage techniques to evade detection. It proposes enhancements to GNN-based detectors to improve their robustness against such sophisticated attacks. The study's findings are crucial for developing more resilient fraud detection systems capable of adapting to evolving fraudulent tactics.

Li et al. [19] introduced the LGM-GNN (Local and Global Aware Memory-Based Graph Neural Network) specifically designed for fraud detection. Their model leverages both local and global graph information to improve detection accuracy. The local module captures the immediate neighborhood information, while the global module aggregates information from distant nodes, ensuring comprehensive pattern recognition in transaction data.

The application of GCNs in anomaly detection extends beyond fraud detection. For instance, the DSTAGCN model (Dynamic Spatial-Temporal Adjacent Graph Convolutional Network) [20] focuses on capturing complex and dynamic dependencies in data. Although primarily applied in traffic forecasting, the principles of dynamic graph representation and spatial-temporal modeling are highly relevant to financial fraud detection, where transaction patterns can also exhibit temporal dependencies.

The integration of GCNs in fraud detection frameworks offers several advantages. GCNs can

capture the intricate relationships between transactions, allowing for more accurate and early detection of fraudulent activities. Additionally, the ability to process large-scale graph data enables these models to handle real-world financial transaction datasets effectively [21].

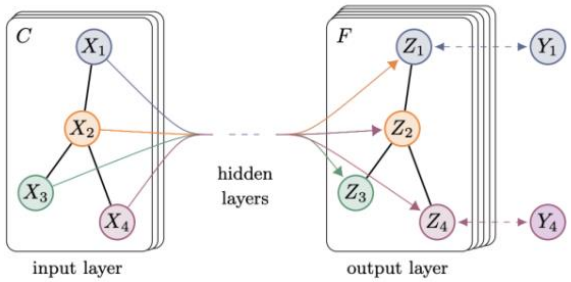

Figure 1. Graph convolution network.

3. Methodology

In this section, we elaborate the data, its preparation and our model for fraud detection.

3.1. Dataset

The dataset used for this study is the **Credit Card Fraud Detection dataset** available on Kaggle [22]. This dataset contains credit card transactions made by European cardholders in September 2013. The dataset presents transactions that occurred over two days, with a total of 284,807 transactions. Among these transactions, 492 are labeled as fraudulent, representing approximately 0.172% of the total transactions. This highly imbalanced nature of the dataset poses a significant challenge for the development of effective fraud detection models. The features are the result of a PCA transformation [23] to protect user identities and sensitive features. These features contain no direct meaning but help in the classification task.

The features were normalized using the Standard Scaler from scikit-learn. Normalization is crucial for ensuring that all features contribute equally to the model's performance, as it scales the features to have a mean of zero and a standard deviation of one.

The distribution of certain features of the dataset is shown in the figure 2. As can be seen, some features contain outliers. In training regular models, this would typically lead us to identify and remove these outliers. However, in this study and with this dataset, our analysis revealed that about half of the outliers belong to the anomalous category. Removing these outliers would further exacerbate our imbalance issue.

3.2. Graph construction

A k-nearest neighbors graph [24] was constructed using the normalized data. In this graph, each transaction is represented as a node, and edges are created based on the similarity between transactions. The similarity is determined using the Euclidean distance metric, and each node is connected to its k nearest neighbors (k=10 in this case). This approach captures the local structure of the data, which is essential for the effectiveness of the Graph Convolutional Network (GCN).

The adjacency matrix generated from the k-nearest neighbors algorithm was converted to edge indices suitable for PyTorch [25] Geometric. The features and labels were then packaged into a Data object, which is the standard input format for PyTorch Geometric models.

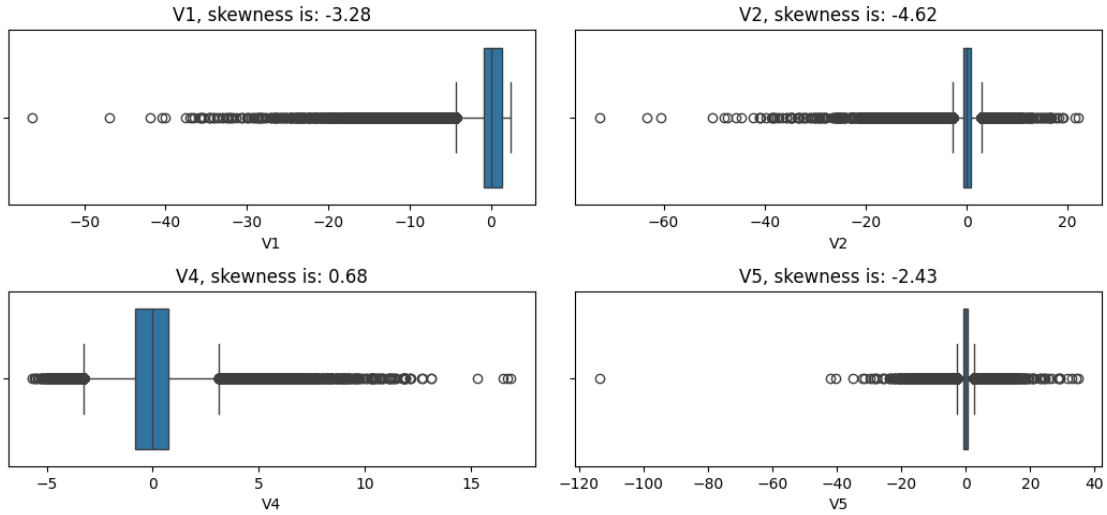

Figure 2. Skewness of some of the dataset features.

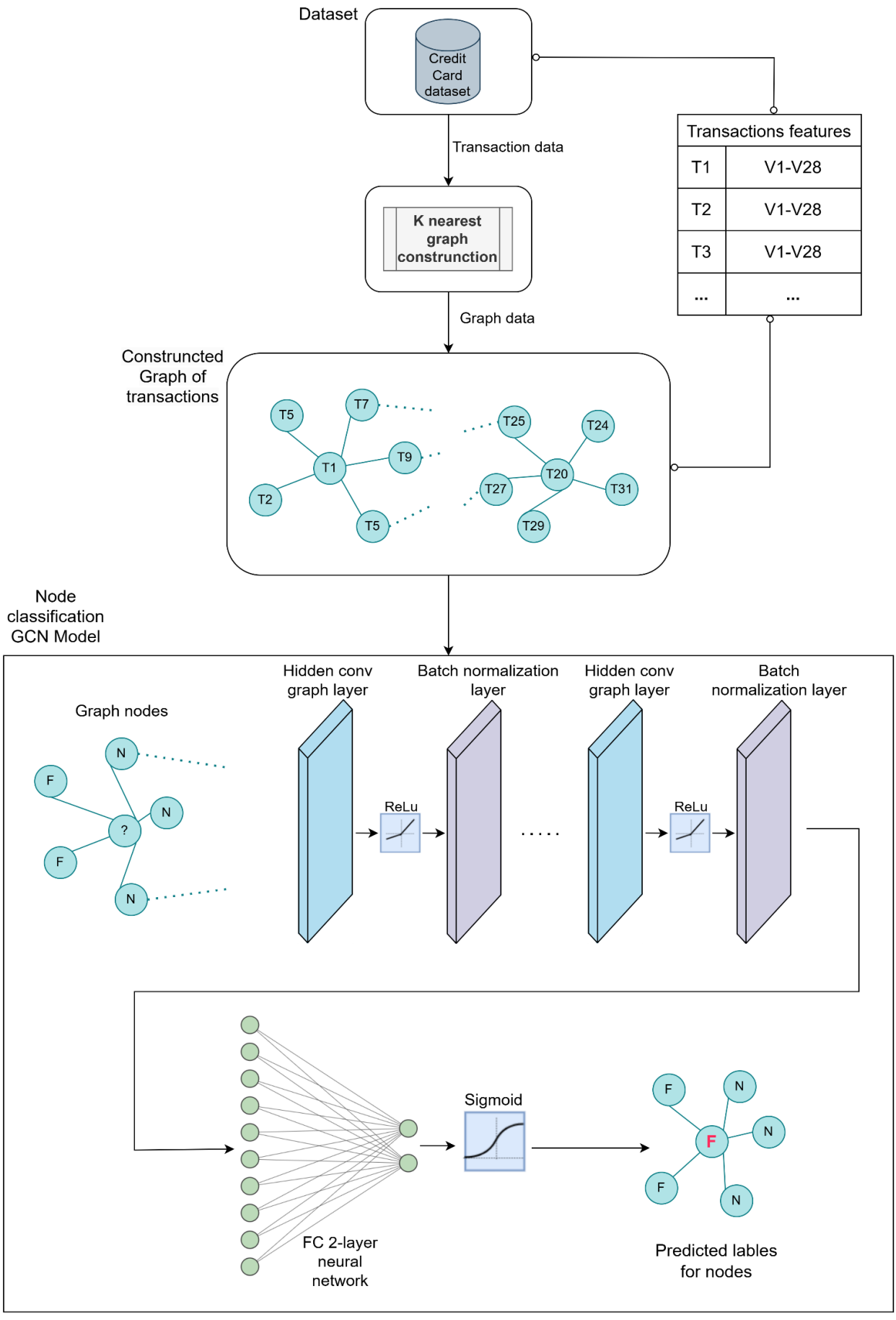

Figure 3. Architecture of the proposed model: FinFD-GCN.

As you can see in figure 5, each node is a transaction that has V1 to V28 features. As the transactions have labels in our dataset, nodes here have labels and thus the task we are doing in FinFD-GCN is node classification.

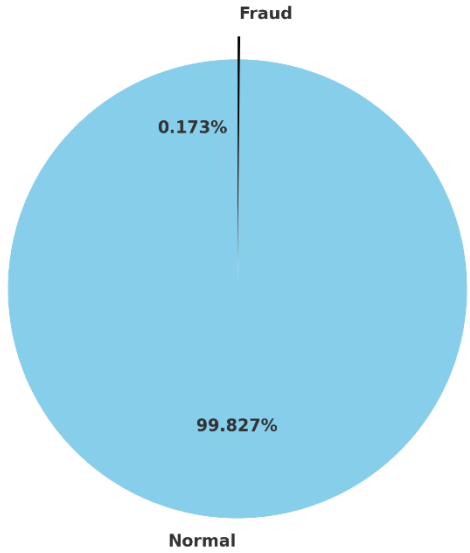

Figure 3. Distribution of the dataset's label which is highly imbalanced.

3.3. Model

The GCN model, FinFD-GCN, designed for this task consists of five graph convolutional layers, each followed by a batch normalization layer. The architecture leverages the ability of GCNs to aggregate information from a node's neighbours to learn meaningful representations for each transaction. The layers progressively transform the input features, enabling the model to capture complex patterns indicative of fraudulent behaviour.

Graph Convolutional Layers: The first layer takes the input features and transforms them to a higher-dimensional space. Subsequent layers further refine these representations by incorporating information from neighbouring nodes. This process allows the model to learn hierarchical feature representations that are crucial for distinguishing between fraudulent and legitimate transactions.

Batch Normalization: Batch normalization [26] is applied after each convolutional layer to stabilize and accelerate the training process. It normalizes the output of the convolutional layers, ensuring that the model trains efficiently.

Dropout Layer: A dropout layer is included before the final classification layer to prevent overfitting. Dropout randomly sets a fraction of the input units to zero during training, which helps the model generalize better to unseen data.

Classification Layer: The final fully connected layer outputs the probability scores for the two classes (fraudulent or legitimate). The model uses the SoftMax function to convert these scores into probabilities.

Training Setup: The model was trained using the Adam optimizer with a learning rate of 0.01. The loss function used was Cross Entropy Loss, which is suitable for multi-class classification tasks. The training process involved updating the model's parameters to minimize the loss function.

Data Splitting: The dataset was split into training and testing sets using a mask-based approach. 80% of the nodes were used for training, while the remaining 20% were reserved for testing. This split ensures that the model's performance is evaluated on unseen data, providing a realistic estimate of its generalization ability.

Training Loop: The model was trained for 100 epochs. In each epoch, the model's parameters were updated based on the training loss, and the model's performance was evaluated on the test set. The training loop included functions to calculate the loss and accuracy, providing feedback on the model's progress throughout the training process.

4. Evaluation

For evaluating the proposed models for classification and sampling, we use common machine learning metrics: F1-score, precision, and recall. These metrics provide different interpretations of the results derived from the confusion matrix, each highlighting various aspects of the model's performance.

Generally, F1-score, being the harmonic mean of precision and recall, is considered an important metric in studies. However, in the context of our discussion, not all these metrics hold the same level of importance. In the dataset under review, which consists of credit card transactions, the critical part of the model is to identify all or most of the anomalous data. Here, anomalous data refers to fraudulent transactions. It is crucial for us to ensure that, as much as possible, all anomalies are detected. Therefore, recall is the most suitable metric for this purpose.

It is advisable to focus our evaluation on improving the recall metric. This metric indicates the proportion of actual anomalies that have been correctly identified. In this case, having a higher number of false positives (incorrectly flagged fraudulent transactions) is less concerning than having a higher number of false negatives (missed fraudulent transactions). The emphasis is on minimizing the number of undetected fraudulent

transactions to ensure comprehensive anomaly detection.

Additionally, we use the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) as an evaluation metric [27]. AUC is a valuable measure because it provides an aggregate performance assessment across all classification thresholds. It represents the likelihood that the model will correctly distinguish between a randomly chosen fraudulent transaction and a randomly chosen legitimate one.

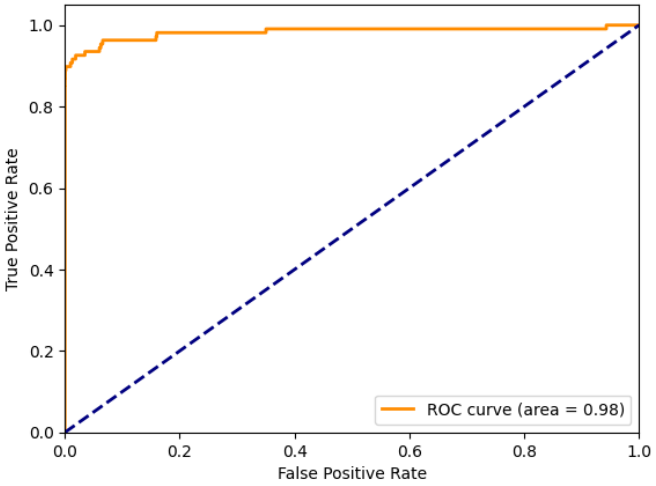

Figure 5. ROC curve of our GCN model.

In the context of using Graph Convolutional Networks (GCNs) for fraud detection in an imbalanced dataset, higher recall and AUC scores are particularly significant. GCNs excel at capturing complex relationships and dependencies within graph-structured data, which is crucial for identifying subtle patterns associated with fraudulent transactions.

A higher recall means the GCN model is successfully identifying a larger proportion of actual fraudulent transactions. This is especially important in an imbalanced dataset where fraudulent transactions are rare compared to legitimate ones. Missing fraudulent transactions can lead to significant financial losses, so maximizing recall ensures that the model captures as many fraudulent activities as possible, even if it means accepting more false positives.

Similarly, a higher AUC score indicates that the GCN model is better at distinguishing between fraudulent and legitimate transactions across all decision thresholds. This metric is crucial because it reflects the model's overall ability to handle the imbalanced nature of the dataset. A high AUC score suggests that the model effectively captures the underlying structure and nuances in the data, leading to better performance in identifying fraudulent transactions.

In summary, higher recall and AUC scores in GCN compared to Random Forest [10], LR [9] and SVM [11] demonstrate the model's capability to catch more intricate details within the dataset, particularly the minority class (fraudulent transactions). This makes GCN a powerful tool for fraud detection, ensuring comprehensive and accurate identification of anomalies in financial data.

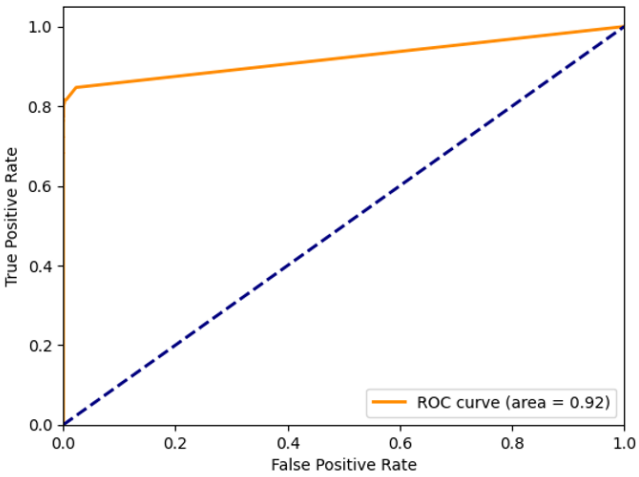

Figure 6. ROC curve of the Random Forest model.

Table 1. Results of FinFD-GCN compared to other machine learning methods.

Model	Evaluation metrics comparison			
	Precision	**Recall**	**F1**	**AUC**
Random Forest	97	73	84	92
Logistic regression	86	52	65	97
SVM	99	68	81	92
FinFD-GCN	92	**81**	**86**	**98**

5. Conclusion

In this study, we explored the application of Graph Convolutional Networks (GCNs) for fraud detection in credit card transaction datasets. Given the imbalanced nature of fraud detection datasets, where fraudulent transactions are rare compared to legitimate ones, traditional machine learning models often struggle to capture the complex relationships and subtle patterns indicative of fraud. GCNs, with their ability to model and analyze graph-structured data, offer a promising solution to this challenge.

We proposed a method called FinFD-GCN which involved several key steps: data preprocessing and normalization, construction of a k-nearest neighbors graph to represent transaction relationships, and conversion of this graph into a format suitable for GCN processing. FinFD-GCN consists of a multi-layer GCN model that leverages batch normalization and dropout layers to enhance learning and prevent overfitting.

The results demonstrated that our GCN model effectively captures the intricate relationships within the transaction data, leading to superior performance in detecting fraudulent transactions. The high recall and AUC-ROC scores underscore the model's ability to identify a significant proportion of actual frauds while maintaining a strong capability to distinguish between fraudulent and legitimate transactions.

These findings highlight the potential of GCNs to address the limitations of traditional methods in fraud detection. By prioritizing recall, we ensured that the model focuses on minimizing false negatives, which is critical in financial contexts where undetected fraudulent transactions can lead to substantial losses. The high AUC-ROC score further confirms the model's robustness in handling the imbalanced nature of the dataset.

FinFD-GCN bridges the gap between traditional and graph-based methods by leveraging graph neural networks to capture intricate transaction patterns, making it an innovative solution for credit card fraud detection.

References

[1] S. prakash. "Online Fraud Is Bad For Business And Customers." Forbes. https://www.forbes.com/sites/forbestechcouncil/2020/07/21/online-fraud-is-bad-for-business-and-customers (accessed).

[2] K. G. Al-Hashedi and P. Magalingam, "Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019," *Computer Science Review*, vol. 40, p. 100402, 2021/05/01/ 2021, doi: https://doi.org/10.1016/j.cosrev.2021.100402.

[3] S. Barman, U. Pal, M. Sarfaraj, B. Biswas, A. Mahata, and P. Mandal, "A complete literature review on financial fraud detection applying data mining techniques," *International Journal of Trust Management in Computing and Communications*, vol. 3, p. 336, 01/01 2016, doi: 10.1504/IJTMCC.2016.084561.

[4] N. Boutaher, A. Elomri, N. Abghour, K. Moussaid, and M. Rida, *A Review of Credit Card Fraud Detection Using Machine Learning Techniques*. 2020, pp. 1-5.

[5] W. Hilal, S. Gadsden, and J. Yawney, "Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances," *Expert Systems with Applications*, vol. 193, p. 116429, 12/01 2021, doi: 10.1016/j.eswa.2021.116429.

[6] A. Trozze *et al.*, "Cryptocurrencies and future financial crime," *Crime Science*, vol. 11, 01/05 2022, doi: 10.1186/s40163-021-00163-8.

[7] M. Nelsen. "Outsmarting Fraudsters with Advanced Analytics." visa. https://usa.visa.com/visa-everywhere/security/outsmarting-fraudsters-with-advanced-analytics.html (accessed).

[8] T. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," 09/09 2016.

[9] A. Mahajan, V. S. Baghel, and R. Jayaraman, "Credit Card Fraud Detection using Logistic Regression with Imbalanced Dataset," in *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, 15-17 March 2023 2023, pp. 339-342.

[10] T. Yiu. "Understanding Random Forest." Towards Data Science. https://towardsdatascience.com/understanding-random-forest-58381e0602d2 (accessed).

[11] N. K. Gyamfi and J. D. Abdulai, "Bank Fraud Detection Using Support Vector Machine," in *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 1-3 Nov. 2018 2018, pp. 37-41, doi: 10.1109/IEMCON.2018.8614994.

[12] S. Beigi and M. R. Amin Naseri, "Credit Card Fraud Detection using Data mining and Statistical Methods," (in en), *Journal of AI and Data Mining*, vol. 8, no. 2, pp. 149-160, 2020, doi: 10.22044/jadm.2019.7506.1894.

[13] P. W. Battaglia *et al.*, "Relational inductive biases, deep learning, and graph networks," *ArXiv*, vol. abs/1806.01261, 2018.

[14] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral Networks and Locally Connected Networks on Graphs," *CoRR*, vol. abs/1312.6203, 2013.

[15] H. Cai, V. W. Zheng, and K. C.-C. Chang, "A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications," *IEEE Transactions on*

Knowledge and Data Engineering, vol. 30, pp. 1616-1637, 2017.

[16] M. Henaff, J. Bruna, and Y. LeCun, "Deep Convolutional Networks on Graph-Structured Data," *ArXiv*, vol. abs/1506.05163, 2015.

[17] X. Ma *et al.*, "A Comprehensive Survey on Graph Anomaly Detection With Deep Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12012-12038, 2023, doi: 10.1109/TKDE.2021.3118815.

[18] Y. Dou, Z. Liu, L. Sun, Y. Deng, H. Peng, and P. S. Yu, "Enhancing Graph Neural Network-based Fraud Detectors against Camouflaged Fraudsters," presented at the Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Virtual Event, Ireland, 2020. [Online]. Available: https://doi.org/10.1145/3340531.3411903.

[19] P. Li, H. Yu, X. Luo, and J. Wu, "LGM-GNN: A Local and Global Aware Memory-Based Graph Neural Network for Fraud Detection," *IEEE Transactions on Big Data*, vol. 9, no. 4, pp. 1116-1127, 2023, doi: 10.1109/TBDATA.2023.3234529.

[20] Q. Zheng and Y. Zhang, "DSTAGCN: Dynamic Spatial-Temporal Adjacent Graph Convolutional Network for Traffic Forecasting," *IEEE Transactions on Big Data*, vol. 9, no. 1, pp. 241-253, 2023, doi: 10.1109/TBDATA.2022.3156366.

[21] Y. Liu *et al.*, "Pick and Choose: A GNN-based Imbalanced Learning Approach for Fraud Detection," presented at the Proceedings of the Web Conference 2021, Ljubljana, Slovenia, 2021. [Online]. Available: https://doi.org/10.1145/3442381.3449989.

[22] M. L. G.-. ULB. "Credit Card Fraud Detection." https://www.kaggle.com/datasets/mlgulb/creditcardfraud (accessed).

[23] H. Zhou, L. Wei, G. Chen, P. Lin, and Y. Lin, *Credit Card Fraud Identification Based on Principal Component Analysis and Improved Adaboost Algorithm*. 2019, pp. 507-510.

[24] WDong, M. Charikar, and K. Li, "Efficient k-nearestnighbor graph construction for generic similarity measures," in *The Web Conference*, 2011.

[25] "Pyhon machine learning library based on torch." https://pytorch.org/ (accessed).

[26] S.Ioffe and C. Szegedy, "Batch normalization: acceleratng deep network training by reducing internal covariate shift," presented at the Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, Lille, France, 2015.

[27] E. ichardson, R. Trevizani, J. A. Greenbaum, H. Carter, . Nielsen, and B. Peters, "The receiver operating characteristic curve accurately assesses imbalanced datasets," *Patterns*, vol. 5, no. 6, p. 100994, 2024/06/14/ 2024, doi: https://doi.org/10.1016/j.patter.2024.100994.

استفاده از شبکه‌های پیچشی گرافی برای تشخیص تقلب در داده‌های مالی

محمد مهدی یادگار و حسین رحمانی*

دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران

ارسال ۲۰۲۴/۰۸/۰۳؛ بازنگری ۲۰۲۴/۱۱/۲۳؛ پذیرش ۲۰۲۴/۱۲/۲۱

چکیده:

در سال‌های اخیر، فناوری‌های نوین تحولات چشمگیری در حوزه مالی و تجاری ایجاد کرده‌اند و همزمان فرصت‌های متعددی را برای کلاه‌برداران فراهم آورده‌اند تا با انجام فعالیت‌های متقلبانه، خسارات کلانی به شرکت‌ها وارد کنند. الگوریتم‌های یادگیری ماشین در سال‌های اخیر به‌طور گسترده‌ای برای تشخیص تقلب در داده‌های مالی به کار گرفته شده‌اند. اما یک چالش رایج، عدم‌توازن در مجموعه داده است که کارایی روش‌های سنتی یادگیری ماشین را محدود می‌کند. یافتن بهترین رویکرد برای مقابله با این عدم‌توازن در داده‌ها، مشکلی است که بسیاری از پژوهشگران هنگام به‌کارگیری روش‌های یادگیری ماشین با آن روبه‌رو هستند. در این مقاله، ما روشی به نام FinFD-GCN پیشنهاد می‌کنیم که از شبکه‌های کانولوشنی گراف برای تشخیص تقلب در مجموعه داده‌های تراکنش‌های کارت اعتباری استفاده می‌کند. در روش FinFD-GCN تراکنش‌ها به صورت یک گراف مدل می‌شوند، به‌طوری که هر گره نمایانگر یک تراکنش و هر یال نشان‌دهنده شباهت میان آن‌ها است. با بهره‌گیری از این نمایش گرافی، FinFD-GCN می‌تواند روابط پیچیده و ناهنجاری‌هایی را کشف کند که در روش‌های سنتی یا نادیده گرفته می شوند و یا حتی امکان شناسایی آن‌ها وجود ندارد. برای ارزیابی روش پیشنهادی، از معیارهای متداول استفاده شده است. نتایج نشان می‌دهد روش FinFD-GCN از نظر شاخص‌های یادآوری و سطح زیر منحنی در مقایسه با روش‌های سنتی نظیر رگرسیون لجستیک، ماشین بردار پشتیبان و جنگل تصادفی بهبود چشمگیری داشته و در زمینه تشخیص تقلب در کارت‌های اعتباری، راهکاری قدرتمند محسوب می‌شود و نسبت به مدل‌های پایه، به‌ترتیب در معیارهای F1 و AUC بهبود ٪۵ و ٪۱۰ حاصل شده است.

کلمات کلیدی: تشخیص تقلب، یادگیری ماشین، کارت اعتباری، شبکه‌های پیچشی گرافی، نمایش گرافی، دسته‌بندی گره.