Sādhanā Vol. 123, No.4, January 2016, pp.2333–2335
DOI 12.3456/s78910-011-012-3

Graph Analysis Augmentation for Enhanced Medicare Fraud Detection: Leveraging Interconnected Data for Superior Accuracy

G. SUGANESHWARI[1,*], C. N. VIGNESHWAR[1], A. SUDHA [1,2] and A. JOTHI[3]

[1] School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, 600127
[2] Center of Smart Grid, Vellore Institute of Technology, Chennai, 600127
[3] Department of Computer Science, Kakatiya Institute of Technology and science, Warangal.

Abstract. Every year, Medicare fraud costs billions of dollars in significant financial losses. The effectiveness of using graph analysis to improve fraud detection in the healthcare industry is examined in this study. By capturing intricate relationships between healthcare entities, our approach aims to improve accuracy and strengthen security protocols. We construct a logical graph structure using diverse datasets, ensuring a comprehensive representation of the healthcare ecosystem. Nodes in the graph correspond to entities (e.g., hospitals, physicians), while edges denote relationships (e.g., referrals, billing patterns). Initially, we employ Random Forest baseline models, achieving an accuracy of 74%. Augmenting these models with graph-based features significantly boosts accuracy to 95%. These features include centrality measures, connectivity patterns, and community detection. Leveraging community detection algorithms, we identify clusters of related entities within the graph. By incorporating community information, our fraud detection accuracy reaches an impressive 99%. Our cooperative method demonstrates the effectiveness of graph analysis in combating fraud. Notably, when combined with traditional baseline models, graph-based approaches significantly enhance detection capabilities. These results underscore the potential of graph analytics to fortify security protocols across various industries, safeguarding against fraudulent activities.

Keywords. Medicare fraud, Graph analysis, Fraud detection, Baseline models, Community detection.

1 Introduction

The evolving landscape of network technologies has engendered a proliferation of sophisticated fraudulent activities across multiple sectors, causing considerable economic detriment. Medicare fraud, for example, costs the European Union around 13 billion euros annually and the United States between 21 and 71 billion dollars [1]. This fraud amounts to around 3% of all Medicare spending and may be above $300 billion yearly. Collusion among multiple parties significantly exacerbates the harm wrought by fraud, underscoring the imperative to scrutinise collusion relationships to bolster fraud detection efforts [2, 3].

Machine learning (ML) approaches have become increasingly popular in the healthcare business as a means of detection and prevention due to the growing problem of Medicare fraud. When examining enormous volumes of Medicare data to find suspect trends or abnormalities suggestive of fraudulent claims, machine learning techniques are essential. Techniques such as supervised, unsupervised, and deep learning are applied to discern anomalies within billing records, patient information, and provider behaviors. Ensemble methods, such as Random Forests, have shown promise in identifying patterns of Medicare fraud, underscoring the significance of ML-based approaches in combatting fraudulent activities within healthcare reimbursements [4]. Furthermore, researchers emphasize the importance of leveraging diverse datasets and innovative algorithms to achieve higher accuracy and adaptability in identifying evolving fraud schemes within Medicare claims [5]. These ML-driven fraud detection systems contribute significantly to mitigating financial losses and upholding the integrity of the healthcare system while ensuring appropriate patient care.

Graph analysis, leveraging network features like centrality measures, has emerged as an influential method for identifying anomalies within medical provider networks. Recent research endeavors delve into integrating graph-related features into traditional ML models to fortify fraud detection systems. This concerted approach seeks to harness insights from network structures, presenting a promising strategy to combat fraudulent activities. Critical for healthcare, the Medicare program grapples with substantial financial losses due to fraudulent activities orchestrated by intricate networks of providers, beneficiaries, and various entities. The complexity of these networks poses a significant challenge to traditional detection methods. Our proposed solution involves an enhanced Medicare fraud detection system that capitalizes on graph analytics approaches. This innovative system aims to proficiently identify potentially fraudulent claims and predict provider fraudulence by employing a combination of graph analytics and ML. Additionally, the system endeavors to uncover hidden relationships among healthcare providers

*For correspondence

and beneficiaries engaged in fraudulent activities. Such an integrated approach is poised to revolutionize fraud detection within the Medicare framework, offering a robust and adaptive system to combat pervasive fraudulent practices.

The Medicare program grapples with substantial financial losses due to these intricate networks of fraudulent activities. Traditional detection methods face challenges in navigating the complexity of these networks. As a response, the proposed solution advocates for an enhanced Medicare fraud detection system that seamlessly integrates graph analytics approaches. This innovative system not only aims to identify potentially fraudulent claims and predict provider fraudulence proficiently but also endeavors to uncover hidden relationships among healthcare providers and beneficiaries engaged in fraudulent activities. Such an integrated approach is poised to revolutionize fraud detection within the Medicare framework, offering a robust and adaptive system to combat fraudulent practices' pervasive and evolving nature. The main contributions of the paper are summarised as follows:

- We propose a Graph-Based Medicare Fraud Detection System to capture interconnected linkages in the graph between beneficiaries, doctors, and medical providers.
- We integrate Community Detection for further improvement by adding a community detection layer within the graph to understand network structures.

In summary, our proposed approach represents a significant advancement in Medicare fraud detection, offering a comprehensive and adaptive system to combat the pervasive and evolving nature of fraudulent practices. By seamlessly integrating graph analytics with machine learning, we aim to revolutionize fraud detection within the Medicare framework, safeguarding the integrity of the healthcare system and ensuring appropriate patient care.

The rest of the paper is organised as follows: Section 2 depicts the literature survey. Section 3 deals with the proposed method. Section 4 concentrates on the experimental results followed by the conclusion.

2 Literature survey

2.1 *Machine Learning (ML) based Approaches*

Several research studies highlight the robustness of ML in detecting and preventing fraudulent activities across various domains. "Credit Card Fraud Detection Using ML" and "Online Fraud Detection Using ML" emphasize the efficiency of ML, especially deep learning algorithms, in identifying fraudulent transactions and activities in credit card usage and online platforms [6, 7, 8]. Similarly, research on "Fraud Detection in Online Advertising" explores ML and network analysis to identify click fraud and bot activities, enhancing fraud detection in digital advertising networks. [9] presented a study, focusing on healthcare fraud detection using data balancing techniques and various ML methods. This research emphasizes addressing data imbalances to optimize fraud detection systems in healthcare [9]. Other studies like "Gradient Boosted Decision Tree Algorithms for Medicare Fraud Detection," "Medical Fraud and Abuse Detection System Based on ML" underscore ML's role in fortifying fraud detection mechanisms across healthcare and diverse sectors [10, 11]. Additionally, comparative analyses like Lavanya et al. (2021) stress different ML approaches in healthcare fraud detection, emphasizing the need for sophisticated and scalable techniques [12]. Óskarsdóttir et al. explore social network analytics for supervised fraud detection in insurance, integrating network-based features to enhance accuracy in detecting potential fraudulent activities within insurance systems [13].

Most existing ML methods become evident due to their narrow focus on modeling users' historical behaviour sequences. These methods often neglect crucial domain knowledge information related to interconnected relationships among medical providers, beneficiaries, and physicians within the healthcare system which leads to the decline in performance of the models. This oversight poses a significant challenge in effectively capturing and understanding the complex network structures inherent in fraudulent activities.

2.2 *Graph-based fraud detection methods*

Research studies demonstrate the effectiveness of graph-based methodologies in diverse fields like fraud detection, credit rating, cybersecurity, and community analysis. "Friend or faux: Graph-based early detection of fake accounts on social networks". It focuses on identifying fake user accounts by establishing a relationship graph of users, evaluating the likelihood of an account being fake based on creation time and responses to friend requests [14]. Studies show graph analysis has improved credit rating models and credit risk assessment by analyzing relationships among borrowing companies [15]. Graph Neural Networks (GNNs) have revolutionized various domains, including credit rating, fraud detection, and recommendation systems. Research by Feng [17] develops credit rating models using GNNs, while Kudo et al. (2020) use Graph Convolutional Networks (GCN) to detect fraud in social networks [16, 17, 18]. Additionally, studies are focusing on specific GNN types like Heterogeneous Graph Neural Networks (HGNNs), Temporal Graph Convolutional Networks (TGCNs), and Hierarchical Attention-based GNNs, demonstrating their efficacy in fraud detection within various domains. Studies also explore the application of GNNs in cybersecurity, corporate credit ratings, online review systems, and financial fraud detection, showcasing the adaptability and robustness of graph-based models [19, 20, 21, 22]. Overall, these research papers highlight the transformative potential of graph-based techniques in understanding complex relationships, detecting anomalies, and improving predictive models across diverse domains.

Most graph-based methods have a major challenge of interpretability, ie when the number of nodes in a graph increases, the graph becomes more extensive and difficult to interpret. To overcome the problem of interpretability and performance, we introduce a groundbreaking solution to ad-

dress this critical problem – a Graph-Based Medicare Fraud Detection System. Our proposed approach aims to overcome the limitations of prevailing methods by acknowledging the importance of capturing interconnected relationships among medical providers, beneficiaries, and physicians within a graph structure.

3 Methodology

Fig. 1. shows the dataset merging process.

3.1 *Dataset*

The study used publicly available data from Kaggle to analyze fraudulent healthcare providers. The data consisted of four subsets: inpatient, outpatient, beneficiary, and potential fraud provider data. As the first step, we combine the inpatient and outpatient data using BeneID and ProviderID as common columns. Tab. 1 shows the description of features in the final data.

3.2 *Approach for fraud detection*

The fusion of graph analytics and ML to detect Medicare fraud involves a unique approach. The process starts by extracting the relationships between providers and physicians from a combined dataset. These relationships are then converted into a bipartite graph, which allows the use of graph centrality information as feature in ML models. The first step is to consolidate and extract relationships from a dataset containing information about healthcare providers, physicians, patients, beneficiaries, claims, attending physicians, and genders. The transformation step represents the relationships in a bipartite network between physicians and providers as nodes and the services they provide as edges. This structure helps visualize and analyze the intricate network between providers and physicians in the healthcare system. Fig. 2 shows the visualization of the proposed model.

3.3 *Graph Construction*

When tabular data are provided, they must be converted into graph-structured data for calculating the graph metrics. Graph representation allows us to capture and analyze the intricate network of connections, where nodes represent individual physicians and edges signify the relationships or interactions between these physicians and healthcare providers[23]. The edges represent the relationship between providers and physicians and the nodes represent the Physicians. Fig. 3 represents this relationship.

3.4 *Graph Metrics*

We derive nodal centrality features from two bipartite graphs: one depicting provider-physician connections and another representing provider-beneficiary relationships. These measures encompass degree centrality, eigenvector centrality, and PageRank, signifying the significance of nodes within the networks. These centrality metrics are then integrated as input features for conventional ML models, enhancing their predictive capabilities base.

3.4.1 *Degree Centrality (Cd)*

$$Cd(vi) = di \tag{1}$$

Where $Cd(vi)$ represents the degree centrality of node vi, and di is the degree of the i^{th} node. This measure counts the number of edges connected to a node, indicating its connectivity within the network.

3.4.2 *Eigenvector Centrality (Ce)*

$$Ce(vi) = \frac{1}{\lambda} 1/ * \sum A_{j,i} * Ce(vj) \tag{2}$$

Where $Cd(vi)$ represents the eigenvector centrality of node vi, Aj, i is the adjacency matrix, λ is an eigenvalue of the adjacency matrix, and Ce(vj) represents the eigenvector centrality of neighboring nodes. Eigenvector centrality considers the number and quality of connections, emphasizing connections to well-connected nodes and indicating a node's influence within the network.

3.4.3 *Page Rank*

PageRank is a centrality measure that strengthens Katz's centrality limits. The number of outgoing edges is used as a normalizing factor for determining each node's impact using PageRank in order to stop overly high relevance from spreading. Here is how PageRank (Cp) is defined:

$$Cp(vi) = \alpha \sum nj = 1Aj.iCp(vj)doutj + \beta \tag{3}$$

where is a constant, β is the bias term that avoids the zero-centrality value, and $doutj$ is the number of outgoing edges (out-degree).

Evaluates the importance of a physician based on the importance of the physicians linked to them. This metric considers both the quantity and quality of connections, assigning higher scores to physicians connected to other highly reputable or influential peers.

3.5 *Community Detection*

Community detection in network analysis involves identifying closely connected groups of nodes in a network. In Medicare fraud detection, community detection is used to find similarities in behavior among healthcare providers or patients. This helps identify potential fraud instances by analyzing the connections within these groups. The Infomap algorithm utilizes information theory to discover the best network modular structures. When applied to Medicare fraud detection, it can uncover complex patterns in provider-patient networks that may indicate fraudulent activities. Infomap achieves this by simulating random walks, dividing the network into groups, and minimizing the information needed to explain node movement. Infomap effectively detects hierarchical and overlapping community structures in complex networks and can handle noise and outliers well [25].

$$L(M) = q \curvearrowright H(P) + \sum i = 1cp \curvearrowright iH(Pi) \tag{4}$$

Where,

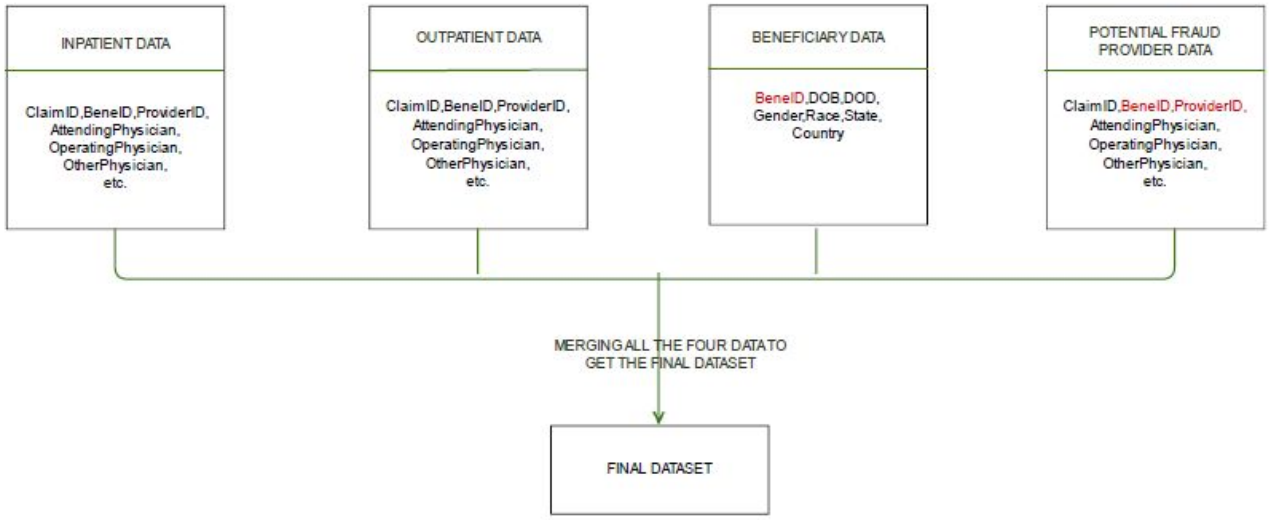

Figure 1. Dataset Merging Process

Table 1. Description of features in the final data

Feature	Description
BeneID	Unique number for beneficiary identification
ClaimID	Unique identification number for claims
ClaimStartDt	Starting date of claim
ClaimEndDt	Ending date of claim
Provider	Unique identification number of the healthcare provider
Attending Physician	Who takes the lead in overseeing the care of a patient
DOB	The date of birth of the beneficiary
DOD	The date of death of the beneficiary
Gender	The gender of the beneficiary
……...	……
PotentialFraud	Potential Fraud labels

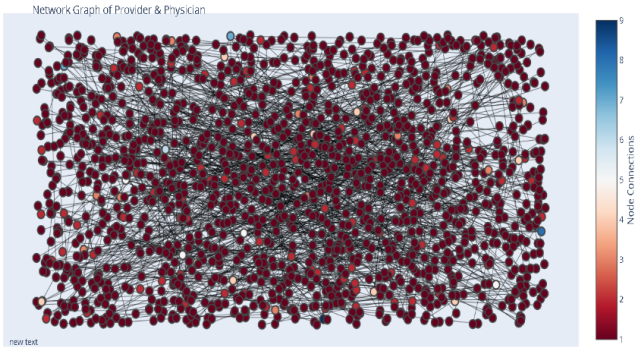

Figure 2. Visualization of Provider-Physician relationship

- $L(M)$ denotes the average description length or code-length of the movement sequences through the network,
- $q \curvearrowright$ represents the probability of choosing a node to move to from the current node,
- $H(P)$ is the entropy of the movement probability distribution over nodes,
- c stands for the number of communities detected,
- $p \curvearrowright i$ signifies the probability of choosing a community to move to, and
- $H(Pi)$ denotes the entropy of the movement probability distribution over communities.

3.6 *ML Models*

Our study employs logistic regression, random forest, and decision tree models—commonly used in fraud detection—to conduct classification tasks for Medicare fraud. The models are evaluated with and without the graph-based features using accuracy and ROC-AUC metrics. Comparing model per-

Table 2. Pseudocode

	InfoMap Algorithm
1:	**Input**:Network G=(V,E),where
	V=set of N vertices,
	E=set of edges
	Minimum quality improvement threshold .
2:	RunPageRanktocalculate vertex visit rate for each vertex.
3:	M={mi={vi}—vi εV}
4:	L=L(M)in(2)
5:	**repeat**
6:	L_{prev}=L
7:	R=random sequence of integers 1 to N
8:	**for** i=0;i¡N;i++do
9:	m_{new}=bestNewModule(M,$vR_[i]$);
10:	Move $vR_[i]$ to m_{new} module, and update M and L.
11:	**end for**
12:	**until** L_{prev} - L < τ
13:	**return** M

Table 3. Performance of the baseline models

S.No	Model	Accuracy
1	Random Forest	74%
2	Decision Tree	71%
3	Logistic Regression	66%

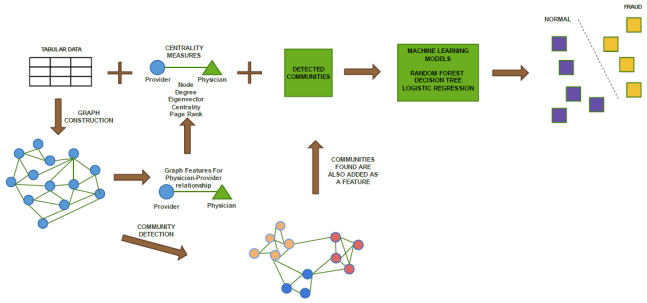

Figure 3. Visualization of model development approach

formances with and without graph-based features helps assess their added value in enhancing Medicare fraud detection accuracy. However, the efficacy of these techniques relies on the quality and relevance of the graph-based features and the complexity of Medicare fraud patterns [26, 27, 28].

4 Results

4.1 *Logistic Regression*

Initially, we chose significant independent variables for constructing a logistic regression model to detect Medicare fraud. The baseline logistic regression model, without node centrality features, showed a lower performance with an accuracy rate of 66

4.2 *Tree-based ML algorithms*

Tree-based models like Random Forests and Decision Trees are important because they are easy to understand, can handle different types of data, and are not prone to overfitting. In a study on Medicare fraud detection, these models were found to perform well compared to logistic regression. They can handle multicollinearity issues when selecting variables, making them suitable for incorporating graph-based features. Adding graph features to these models, especially Random Forest, led to significant improvements in performance. The Random Forest model achieved an impressive 74% accuracy without any tuning of its hyperparameters, while the Decision Tree model had a slightly lower accuracy of 71%. Tables 2 and 3 represent the accuracy of the various baseline models with and without Graph analysis. Fig. 4 shows the accuracy of random forest with baseline model, graph analysis and community detection.

The addition of graph features greatly improved the performance of the Random Forest model in detecting Medicare fraud. This led to a significant increase in accuracy to 95%. This improvement highlights the importance of using graph-based features in Medicare fraud detection. The success of the model was enhanced by incorporating the InfoMap algorithm for community detection. This led to a significant increase in accuracy, from 95% to 99%. This improvement demonstrates the effectiveness of combining graph features with community detection methods like InfoMap in identifying fraud in the Medicare system.

4.3 *Feature Importance*

In Fig 5, we notice that the graph-derived features hold greater importance, displaying higher values compared to the base features. Notably, the feature ”Attending_Physician_cluster” ranks the highest, signifying its utmost importance. Following closely are ”Provider_eigenvector_centrality,” ”Attending _Physician_pagerank,” ”Provider_pagerank,” ”provider_eigenvector _centrality,” and ”provider_degree.” The rest of the graph features exhibit descending importance thereafter.

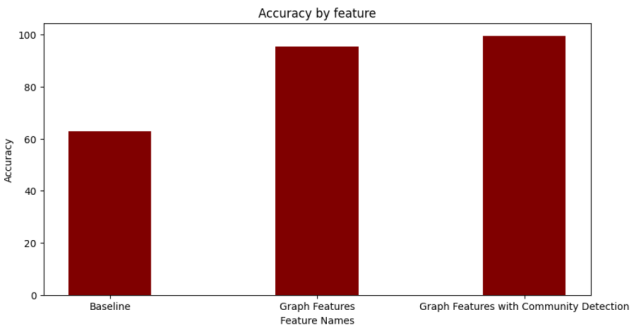

Figure 4. Graph showing the performance of random forest.

Table 4. Performance of model with graph features and with community detection.

S.No	Model	Accuracy
1	Baseline Random Forest	63.18%
2	With Graph Features	95.76%
3	With Community Detection(InfoMap)	99.5%

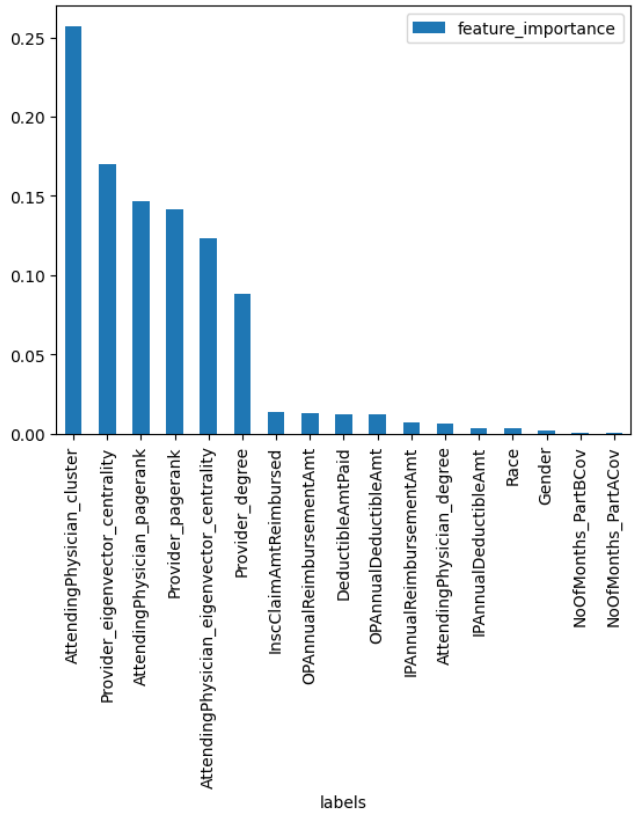

Figure 5. Feature Importance Graph

4.4 *AUC-ROC*

An AUC of 0.558, while indicating some discriminatory ability, suggests that the model's performance might be only slightly better than random chance in distinguishing between the two classes.

5 Conclusion

This study focused on enhancing fraud detection in healthcare using graph analysis. By incorporating graph centrality features derived from provider-physician networks, we improved the accuracy of ML models. Nodal centrality measures like degree centrality, eigenvector centrality, and PageRank were used as input features for Logistic Regression, Random Forest, and Decision Tree models to detect fraudulent medical providers. The application of graph-based learning significantly enhanced the fraud detection model's performance compared to the baseline. However, addressing class imbalance in future studies will be crucial in developing a more universally applicable fraud detection model.

References

[1] Yoo, Yeeun, Jinho Shin, and Sunghyon Kyeong. *Medicare Fraud Detection using Graph Analysis: A Comparative Study of Machine Learning and Graph Neural Networks*. IEEE Access, 2023.

[2] Bartsiotas, George A., and Gopinathan Achamkulangare *Fraud prevention, detection and response in united nations system organizations*. Jenewa: United Nations, 2016.

[3] Kamal, Rabah, and Cynthia Cox. *How has US spending on healthcare changed over time*. Health Spending, 2019.

[4] Emran Nabrawi, and Abdullah Alanazi. *Fraud Detection in Healthcare Insurance Claims Using Machine Learning*. Risks, 2023.

[5] Leelakumar Raja Lekkala. *Importance of Machine Learning Models in Healthcare Fraud Detection*. Voice of the Publisher, 9, 207-215, 2023.

[6] Dornadula, Vaishnavi Nath, and Sa Geetha *Credit card fraud detection using machine learning algorithms*. Procedia computer science, 165, 2019.

[7] Sellam, V, Tushar P, Rohit G. and Sanyam S. *Credit card fraud detection using machine learning. Indian Journal of Computer Graphics and Multimedia*. 1(1), 2021.

[8] Viswanatha, V, Ramachandra A C, Deeksha V, and Ranjitha R. *Online Fraud Detection Using Machine Learning Approach. International Journal of Engineering and Management Research*. 13(4), 2023.

[9] Agrawal, Nikita, and Suvasini Panigrahi. *A Comparative Analysis of Fraud Detection in Healthcare using Data Balancing and Machine Learning Techniques. International Conference on Communication, Circuits, and Systems (IC3S)*. IEEE, 2023.

[10] Hancock, John T, and Taghi M. Khoshgoftaar. *Gradient boosted decision tree algorithms for medicare fraud detection. SN Computer Science 2.4*.2021

[11] Zhang, Conghai, Xinyao Xiao, and Chao Wu *Medical fraud and abuse detection system based on machine learning. International journal of environmental research and public health*. 2020

[12] Lavanya, S., S. Manoj Kumar, and P. Mohan Kumar. *Machine learning based approaches for healthcare fraud detection: A comparative analysis. Annals of the Romanian Society for Cell Biology*. 2021, 8644-8654.

[13] Óskarsdóttir, María, Waqas Ahmed, Katrien Antonio, Bart Baesens, Rémi Dendievel, Tom Donas, and Tom Reynkens. *Social network analytics for supervised fraud detection in insurance. Risk Analysis*. 2022, 1872-1890.

[14] Breuer, Adam, Roee Eilat, and Udi Weinsberg. *Friend or faux: Graph-based early detection of fake accounts on social networks. Proceedings of The Web Conference*. 2020.

[15] Prusti, Debachudamani, Daisy Das, and Santanu Kumar Rath. *Credit card fraud detection technique by applying graph database model. Arabian Journal for Science and Engineering*. 46.9, 2021, 1-20.

[16] Kudo W., Nishiguchi M. and Toriumi F. *GCNEXT: graph convolutional network with expanded balance theory for fraudulent user detection*. Soc. Netw. Anal. Min. 10, 85 (2020).

[17] Feng, Bojing, Haonan Xu, Wenfang Xue, and Bindang Xue. *Every Corporation Owns Its Structure: Corporate Credit Rating via Graph Neural Networks. In Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Cham: Springer International Publishing*. 2022, 688-699.

[18] Li, Ao, Zhou Qin, Runshi Liu, Yiqun Yang, and Dong Li. *Spam review detection with graph convolutional networks. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2703-2711. 2019.

[19] Cheng, Dawei, Xiaoyang Wang, Ying Zhang, and Liqing Zhang. *Graph neural network for fraud detection via spatial-temporal attention. IEEE Transactions on Knowledge and Data Engineering 34*. 3800-3813. 2020.

[20] Ma, Shuai, Jian-Wei Liu, Xin Zuo, and Wei-Min Li. *Heterogeneous graph gated attention network. In 2021 International Joint Conference on Neural Networks (IJCNN)*. 1-6. IEEE, 2021.

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

[22] Liu, Yajing, Zhengya Sun, and Wensheng Zhang. *Improving fraud detection via hierarchical attention-based Graph Neural Network. Journal of Information Security and Applications*. 72. 2023.

[23] Branting L. K., Reeder F., Gold J. and Champney T. Graph analytics for healthcare fraud risk estimation,IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM). pp. 845-851, 2016.

[24] Bauder R. A., and Khoshgoftaar T. M The detection of medicare fraud using machine learning methods with excluded provider labels. In The Thirty-First International Flairs Conference..AAAI,2018.

[25] Smiljanić J., Blöcker C., Holmgren A., Edler D., Neuman M., and Rosvall M. Community detection with the map equation and infomap: Theory and applications.arXiv preprint arXiv:2311.04036, 2023.

[26] Gupta R. Y., Mudigonda S. S. and Baruah P. K. A comparative study of using various machine learning and deep learning-based fraud detection models for universal health coverage schemes. International Journal of Engineering Trends and Technology, 69(3), 96-102,2021.

[27] Bauder R. A., Khoshgoftaar T. M. Medicare fraud detection using machine learning method s. In 2017 16th IEEE international conference on machine learning and applications (ICMLA) pp. 858-865. IEEE,2017.

[28] Hancock J. T., Bauder R. A., Wang H., and Khoshgoftaar T. M. Explainable machine learning models for Medicare fraud detection. Journal of Big Data, 10(1), 154,2023.