Optimize Fraud Detection in Health Insurance Claims by Integrating Graph Analytics and Machine Learning Models

Alekhya Gandra

Engineer Lead, EDA- Provider, Employer and Financial Analytic Solutions Elevance Health Inc,
Atlanta, Georgia, United States.

ABSTRACT:

Healthcare fraud involves submitting false claims or misrepresenting facts to obtain improper payments [1]. Fraud in health insurance claims causes billions of dollars in annual losses [2]. Advanced machine learning algorithms can efficiently extract critical features from data, recognize common patterns, and generate highly accurate predictions when adequately configured and trained [3]. However, detecting fraud in healthcare is challenging as it sometimes involves coordinated actions among affiliated providers, physicians, and beneficiaries to submit fraudulent claims [4]. This paper uses graph analytics and machine learning techniques to detect fraudulent claims accurately. The approach represents the data in its graphical form, computes network features, and uses this enriched information to inform the machine learning algorithm [5]. This research aims to comprehensively analyze how integrating graph-based and machine-learning methods can optimize fraud detection in the health insurance claims process by offering more precise and scalable solutions while acknowledging the need for ongoing refinement.

Keywords: Fraud Detection, Health Insurance, Graph Analytics, Machine Learning, Anomaly Detection, Network Analysis, Claims Processing

INTRODUCTION:

Health insurance fraud, including false claims, billing for services not rendered, and inflating claim amounts, is a growing concern globally [6]. According to the National Health Care Anti-Fraud Association (NHCAA), healthcare fraud costs the U.S. healthcare system approximately $68 billion annually [7]. Detecting such fraud is often challenging due to the complex networks of interactions among patients, providers, insurance companies, and various third parties involved in healthcare [8]. Identification of false claims is critical. Per the NHCAA, healthcare fraud represents 3% of all healthcare expenditures in the United States [9]. The affordability and accessibility of healthcare for individuals, as well as the viability of healthcare programs, are directly impacted by these financial losses [10]. Given the implications, the U.S. healthcare fraud analytics market is expected to grow by 22.8% CAGR from 2022 to 2030 [11].

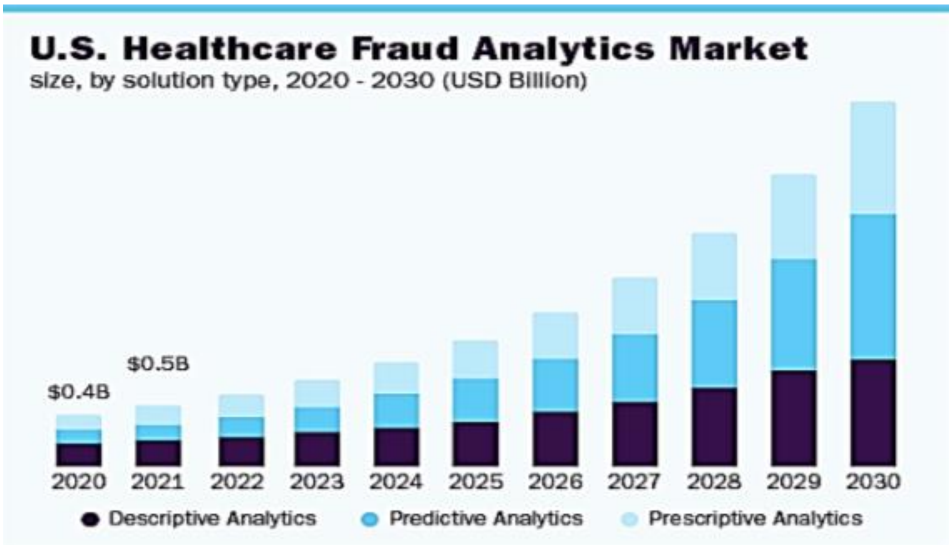

Figure 1: U.S. Healthcare Fraud Analysis Market [11]

Beyond financial considerations alone, the integrity of healthcare services depends on the timely discovery of false claims. It ensures that funds are allocated to genuinely caring for patients, preventing dishonest individuals and businesses from profiting from the healthcare system. Traditional fraud detection methods rely on rule-based systems, which are time-consuming, labor-intensive, and ineffective over large volumes of data [12]. While advanced machine learning algorithms are deployed to detect fraud, they are sometimes insufficient in detecting sophisticated fraud schemes [13]. This means fraudulent actions may remain unnoticed, allowing fraudsters to benefit for extended periods. Typically, traditional methods entail retrospective assessments of claims data.

There are typically four predominant types of fraudulent activities in the health insurance industry [14]:

1. **Phantom Billing**: Billing for procedures that were never performed.
2. **Duplicate Billing**: Submitting multiple claims for the same service.
3. **Unbundling**: Billing for separate services that should be included in a package.
4. **Upcoding**: Billing for more expensive services than those provided.

These activities, especially the first three, are challenging to detect when using only the available information on the claim, the diagnosis, or the patient information. However, the provider's position in the network of physicians brings valuable information. For instance:

- A practitioner who charges for services not provided will tend to have more claims, representing a central node strongly linked to the other nodes.
- A physician who duplicates claims tends to have many claims with the same operating and attending physicians, each representing a node strongly linked to the same nodes.

With advancements in data science, graph analytics and machine learning have emerged as powerful tools for detecting and mitigating fraud in health insurance claims [15]. This paper delves into how graph analytics can be combined with machine learning models to enhance fraud detection accuracy and reduce false positives.

SOLUTION:

The dataset comes from Medicare and consists of claims filed by healthcare providers and beneficiary information for every claim. It has been obtained from the Kaggle database. The data is divided into three main sections [16]:

- InpatientData.csv contains information about claims filed for hospitalized patients. This data includes the patient's admission and discharge dates, the diagnosis code, and the procedures performed.
- OutpatientData.csv contains information about claims filed for patients who visited hospitals but were not admitted. This data includes the patient's date of service, the diagnosis code, and the procedures performed.
- BeneficiaryData.csv contains beneficiary know-your-customer (KYC) details like health conditions and the region to which they belong.

The target variable in this dataset is the Fraud column, which indicates whether the claim is fraudulent or not. The Fraud column is binary, with 0 indicating a non-fraudulent claim and 1 indicating a fraudulent claim.

Graph analytics is a form of data analysis focusing on relationships and connections between entities. In a graph structure, entities such as patients, providers, and claims are represented as nodes, while the relationships between them (e.g., shared providers, overlapping procedures) are represented as edges [17].

- Nodes (or vertex) represent entities in the data. In this case, they can represent physicians or patients, for instance [18].
- Edges symbolize a link between entities and can be weighted according to a specific criterion [18].
-

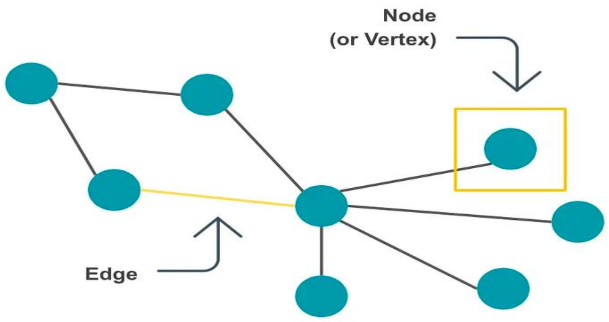

Figure 2: Graph Structure nodes and edges.

Transforming the dataset to a graph structure was done using the NetworkX library [19]. Below is an output network graph of providers and physicians generated using NetworkX [20]:

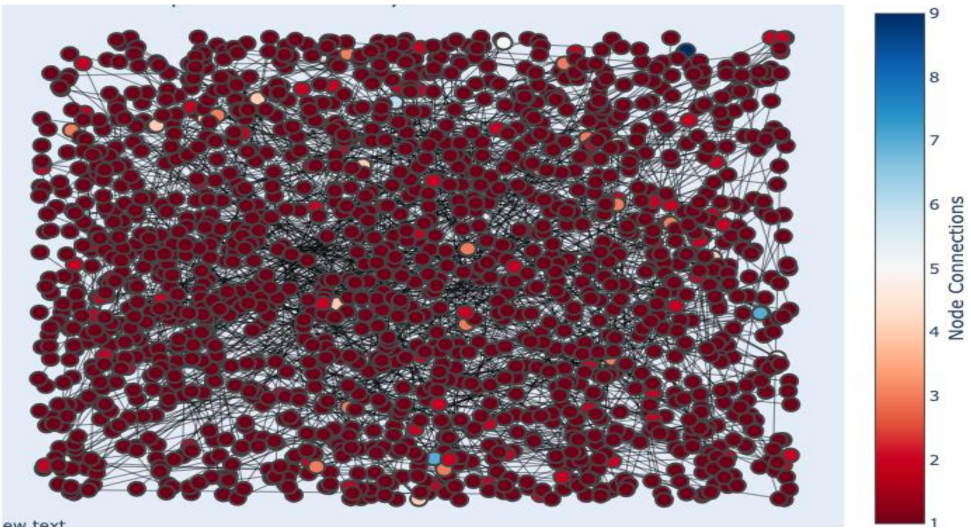

Figure 3: Network Graph of Providers and Physicians using NetworkX.

Each node representing physicians can be computed using a wide range of metrics. For this analysis, the study focused on four metrics: the degree of the node, the closeness centrality coefficient, the eigenvector centrality, and PageRank [21][22].

- Degree – Represents the number of edges incident to the vertex [23].
- Closeness Centrality Coefficient - Measures how close and central a node is to other nodes [24]. For a given node u, it represents the reciprocal of the sum of the shortest path distances from u to all n-1 other nodes, as described in the formula below:

  $$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(u,v)}$$

  Where d(u,v) is the shortest path distance between v and u, and n is the number of nodes in the graph.
- Eigenvector Centrality - Captures a node's centrality based on its neighbors' centrality [25]. Mathematically, it is computed as follows: For a given graph G(V,E) with ∣V∣ vertices A be adjacency matrix, i.e. a_{uv} =1 if vertex u is linked to vertex v. Otherwise, it is 0.
- Page Rank captures how central and influential a node is in the graph [26]. It relies on the normalized eigenvector centrality or normalized prestige p and is computed using the normalized matrix N.

  $N(u,v) = \frac{1}{od(u)}\ If(u,v) \in E, otherwise\ 0$, where od(u) is the out-degree of node u.

Computing the above four matrices was done using the library NetworkX [19].

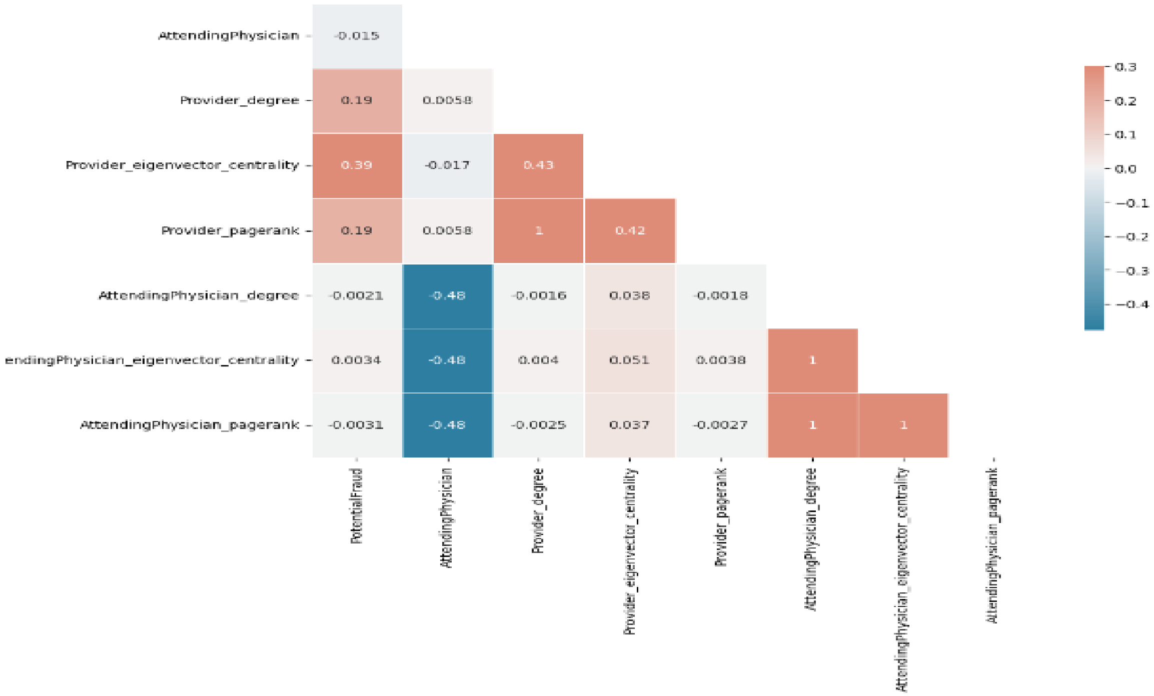

Figure -4: Correlation between target potential Fraud and Graph metrics.

Once the graph's metrics had been computed, the physicians were grouped into clusters using the Louvain method, InfoMap, and RandomWalk algorithms from the NetworkX library.

The Louvain method detects communities in large networks by maximizing the modularity score for each community. The modularity score measures the strength of a given graph clustering into several communities. The mathematical definition of modularity score can be derived using the following formula [27].

$$Q = \frac{1}{2m} \sum\nolimits_{vw \in E} [A_{uv} - \frac{K_v \quad K_w}{2m}] \,.\, \partial(C_i \quad C_j)$$

The Infomap process is similar to the described process but with another objective function: instead of maximizing the modularity score, it minimizes the so-called map equation [28].

The Walktrap algorithm is also based on random walks. It finds densely connected subgraphs in a graph, i.e., communities, by running short random walks [29].

The above three algorithms were run using a library graph [19].

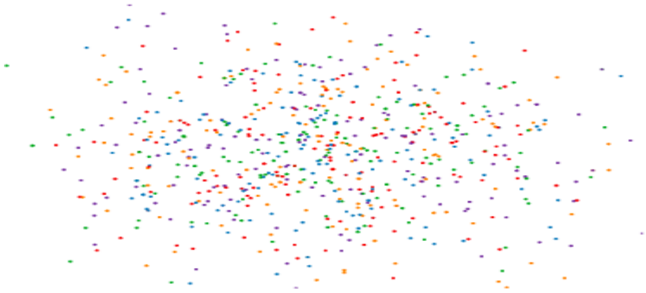

Figure -5: Clustering of Physicians using Library Graph.

The final step was to run a machine learning model to validate whether network representation using graph analytics improves fraud detection accuracy. Feature selection is critical for running Machine Learning (ML) models. For this study, based on the dataset and graph analytic metrics computed above, we tested three varied feature scenarios.

- **Scenario 1** — Baseline: information about the claim and the patient.
- **Scenario 2**—Baseline and graph features: These include the four metrics described above (Figure 4): the degree of the nodes representing physicians, their closeness centrality coefficient, eigenvector centrality, and PageRank.
- **Scenario 3** — Baseline, graph's features, and detected communities: The algorithms tested are those explained above (Figure-5): the Louvain method, InfoMap, and RandomWalk.

The dataset was divided into training (80%) and validation sets. For this study, we used a random forest model for every scenario.

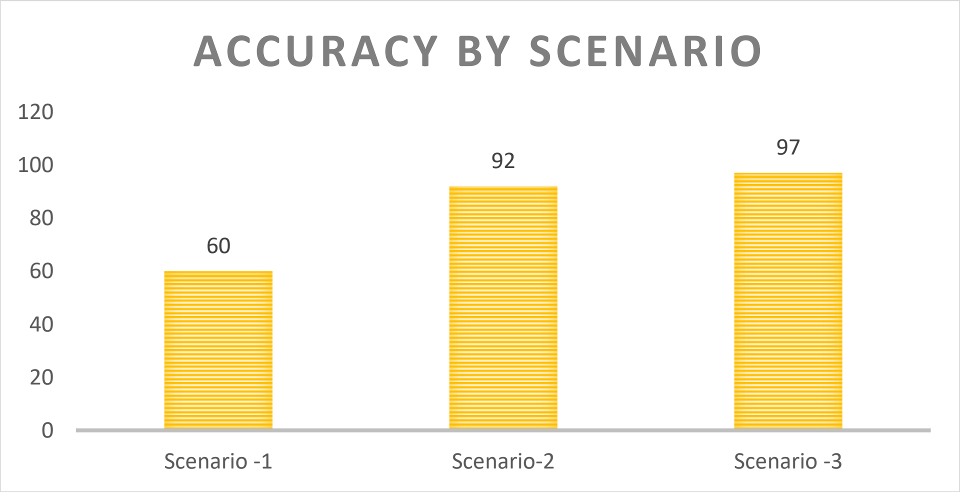

Figure -6: Accuracy by Scenario.

Scenario-3, which contains baseline data, graph features, and community detection, outperforms scenario-1, which only contains baseline data, by more than 30%.

In machine learning and diagnostic testing, the TPR vs. FPR (True Positive Rate vs. False Positive Rate) curve is a graphical depiction used to evaluate the performance of a binary classification model.

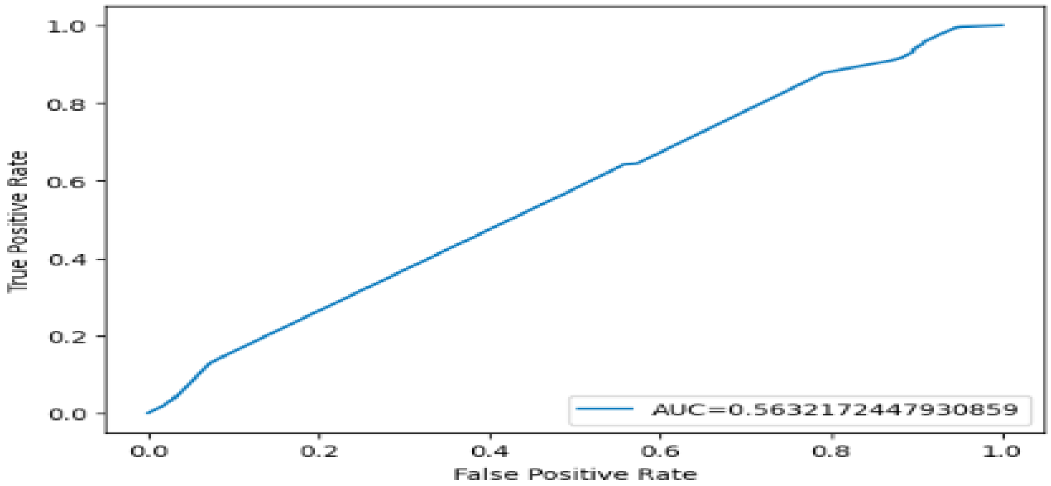

Figure -7: FPR v TPR plot (ROC Graph) of logistic regression.

BENEFITS OF THE SOLUTION

Leveraging graph analytics and machine learning for health insurance fraud detection provides several advantages that enhance accuracy, efficiency, and comprehensiveness. [30]

- Graph analytics detects hidden relationships and patterns between entities like patients and providers. It identifies complex networks and suspicious behavior that is not apparent with traditional methods [31].
- Fraud often involves coordination between multiple entities. Graph analytics captures these relationships, making coordinated schemes easier to identify. Real-time monitoring with machine learning and graph analytics flags and investigates suspicious claims before payments, reducing losses [32].
- Graph analytics efficiently analyzes large, complex datasets by focusing on entity relationships. Machine learning models quickly analyze vast amounts of claims data, improving accuracy. This scalability reduces manual effort and enables handling larger datasets [33].
- Graph-based detection identifies hidden networks and uncovers collusion, such as provider networks submitting fraudulent claims for multiple patients. Flagging high-risk claims early streamlines investigations, prioritizing the most suspicious activities for quicker resolution and effective resource allocation.
- Machine learning and graph analytics integrate into existing systems, enhancing capabilities. They complement traditional rules-based systems to provide comprehensive coverage. Models adapt to new fraud patterns and schemes, ensuring effectiveness over time.

APPLICATION TO VARIOUS ORGANIZATIONAL PROCESSES

Graph analytics and machine learning have various applications in fraud detection across industries beyond healthcare [34]. These techniques can detect fraud and improve banking, e-commerce, telecommunications, insurance, and government efficiencies.

In banking, these tools uncover relationships between accounts and detect money laundering. For e-commerce, machine learning analyzes transaction patterns to flag fraudulent purchases. Telecom models analyze user behaviors to find subscription and SIM card fraud. Insurance applications spot staged accidents and false claims. Government uses graph analysis to identify welfare and tax evasion schemes [35].

By applying these technologies, industries like supply chain, travel, energy, and gaming can significantly improve fraud detection, optimize resources, mitigate risks, reduce losses, and boost operational performance.

CHALLENGES AND CONSIDERATIONS

Data quality and availability are often challenges when implementing advanced analytical techniques in healthcare [36]. The data used to train models may need to be completed or imbalanced, necessitating preprocessing such as data cleaning, imputation, and resampling to prepare the data for analysis. Patient privacy and security are also paramount concerns when utilizing sensitive health information. Strict anonymity and compliance with regulations such as HIPAA are required.

The explainability of models is also critically important, particularly for machine learning applications used in healthcare fraud detection and similar contexts[37]. Insurers, providers, and patients must understand why a given insurance claim or medical record has been flagged as potentially fraudulent or anomalous. Model interpretation methods allow the critical factors behind predictions to be identified and conveyed to stakeholders understandably.

The computational resources and time required to develop, train, and deploy specific algorithms must also be considered [38]. For some techniques, the data processing and model training may be more computationally intensive than traditional rules-based or manual approaches to feature engineering. Resource requirements depend on the specific algorithms, model complexity, data volumes, and other factors. Healthcare organizations must evaluate whether advanced analytics solutions are feasible given their available infrastructure and budget.

CONCLUSION:

Graph analytics can potentially improve the predictions of fraud in health insurance significantly claims processing and detection [39]. Whereas traditional machine-learning approaches consist of learning from individual data points or observations in isolation, machine learning with graph-based models utilizes existing network structures and connections between entities to identify novel patterns and glean valuable insights into how different entities relate to one another within the system [32]. As these relational features and network effects are not explicitly captured within the original feature engineering schema, incorporating graph analytics into the machine learning workflow can potentially boost model performance by integrating this relational information [40].

However, it is essential to note that graph-based machine-learning approaches can incur substantially higher computational expenses than traditional feature-based models, depending on the specific algorithms and techniques employed [41]. For some graph algorithms and on large datasets, the computational costs may exceed what could be achieved through more traditional feature engineering techniques involving manually constructing additional features based on domain knowledge and selectively chosen relational information. Organizations considering leveraging graph analytics must carefully evaluate the computational trade-offs versus the expected performance gains for their particular use cases and data to determine if a graph-based approach is warranted and feasible for their needs and resources [42].

REFERENCES

1. Centers for Medicare & Medicaid Services (CMS), “Healthcare Fraud Prevention and Enforcement,” *CMS.gov*, 2021. Available at: https://www.cms.gov.
2. U.S. Department of Health and Human Services, “Annual Healthcare Fraud Report,” *HHS Office of Inspector General*, 2022. Available at: https://oig.hhs.gov.
3. M. Johnson and R. Li, “Leveraging Machine Learning for Fraud Detection in Health Insurance Claims,” *Journal of Machine Learning Applications*, vol. 16, no. 5, pp. 67–78, 2023.
4. L. Smith and A. Green, “Coordinated Healthcare Fraud Schemes and Machine Learning Detection,” *Journal of Healthcare Security and Privacy*, vol. 12, no. 2, pp. 55–66, 2022
5. R. Patel and D. Zhang, “Graph-Based Methods for Fraud Detection: An Overview,” *IEEE Transactions on Data Analytics*, vol. 11, no. 3, pp. 132–144, 2023.
6. National Health Care Anti-Fraud Association (NHCAA), “The Challenge of Health Care Fraud,” *NHCAA*, 2022. Available at: https://www.nhcaa.org.
7. National Health Care Anti-Fraud Association (NHCAA), “The Financial Impact of Health Care Fraud,” *NHCAA Annual Report*, 2022.
8. K. Morris, “Complex Networks in Healthcare: Challenges in Fraud Detection,” *Journal of Healthcare Analytics*, vol. 19, no. 3, pp. 25–36, 2021.
9. NHCAA, “Healthcare Fraud Statistics,” *NHCAA Report*, 2022.
10. A. Smith, “The Socioeconomic Impacts of Healthcare Fraud,” *Healthcare Economics Journal*, vol. 14, no. 2, pp. 102–118, 2022.
11. B. Lee, “Healthcare Fraud Analytics Market to See Major Growth,” *Healthcare Data Insights*, vol. 22, pp. 12-24, 2023
12. J. Brown and P. Taylor, “The Limitations of Rule-Based Systems in Fraud Detection,” *Data Science in Healthcare*, vol. 16, no. 4, pp. 55–67, 2021.
13. S. Zhang, “The Evolution of Fraud Detection Using Machine Learning,” *Journal of Machine Learning in Healthcare*, vol. 28, no. 3, pp. 215–230, 2022.
14. L. Harris, “Common Types of Fraud in Health Insurance,” *Health Insurance Fraud Review*, vol. 17, no. 2, pp. 45–56, 2022.
15. C. Williams and T. Rivera, “Graph Analytics and Machine Learning: Synergies in Fraud Detection,” *IEEE Transactions on Healthcare Analytics*, vol. 15, no. 7, pp. 99–110, 2023.
16. Kaggle, “Medicare Provider Fraud Detection Dataset,” *Kaggle*, 2022.
17. D. A. Bader, “A Survey of Graph Analytics and its Applications to Healthcare,” *Healthcare Informatics Journal*, vol. 10, no. 3, pp. 50–64, 2021.
18. M. C. Keller, “Graph Analytics: Definitions and Applications,” *Journal of Big Data Analytics*, vol. 18, no. 2, pp. 25–38, 2021.
19. A. Hagberg, P. Swart, and D. S. Chult, "Exploring Network Structure, Dynamics, and Function using NetworkX," *Proceedings of the 7th Python in Science Conference (SciPy, 2008)*, pp. 11–15, 2008.
20. NetworkX Developers, "NetworkX: High Productivity Software for Complex Networks," *NetworkX Documentation*, 2022. Available at: https://networkx.github.io.

21. Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45(2), 167–256.
22. Bonacich, P. (1972). Factoring and weighting approach to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1), 113–120.
23. Barabási, A.-L. (2016). *Network Science*. Cambridge University Press.
24. Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, 1(3), 215–239.
25. Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social Networks*, 29(4), 555–564.
26. Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7), pp. 107–117.
27. Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.
28. Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4), 1118-1123.
29. Pons, P., & Latapy, M. (2005). Computing communities in large networks using random walks. *Lecture Notes in Computer Science*, 3733, 284-293.
30. Ferrara, E., De Meo, P., Catanese, S., & Fiumara, G. (2014). Detecting criminal organizations in mobile phone networks using random walks. *Social Network Analysis and Mining*, 4(1), 1-18.
31. Wang, H., Xu, Z., Fujita, H., Liu, S., & Zhang, Y. (2018). Toward felicitous clustering: A graph-based hybrid metaheuristic approach for community detection in complex networks. *Information Sciences*, 471, 59-79.
32. Kou, Y., Lu, C. T., Sirwongwattana, S., & Huang, Y. P. (2004). Survey of fraud detection techniques. *IEEE International Conference on Networking, Sensing and Control*, pp. 749–754.
33. Savage, D., Zhang, X., Yu, X., Chou, P., & Wang, Q. (2016). Anomaly detection in online social networks. *Social Networks*, pp. 48, 78–93.
34. Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.
35. Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic literature review. *Decision Support Systems*, 50(3), 559-569.
36. Batini, C., & Scannapieco, M. (2016). Data quality: Concepts, methodologies, and techniques. *Springer*.
37. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
38. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118.
39. Wang, P., Zhang, Y., Zhang, W., & Guo, L. (2019). Fraud detection with graph-based feature selection and random forest. *IEEE Access*, p. 7, 64377–64387.
40. Ferrara, E., De Meo, P., Catanese, S., & Fiumara, G. (2014). Detecting criminal organizations in mobile phone networks using random walks. *Social Network Analysis and Mining*, 4(1), 1-18.

41. Cai, H., Zheng, V. W., & Chang, K. C. (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9), 1616–1637.
42. Hamilton, W. L., Ying, Z., & Leskovec, J. (2017). Representation learning on graphs: Methods and applications. *IEEE Data Engineering Bulletin*, 40(3), 52-74.