iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.79131

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

An Explainable Graph-Augmented XGBoost Framework for Health Insurance Fraud Detection

M. Aruna[1], Heerekar Sairaj[2], Vuppala Sahil[3], Thaneeru Harshini[4]
[1]*Assistant Professor, Department of Computer Science and Engineering, Methodist College of Engineering and Technology, Hyderabad, India*
[2, 3, 4]*Students, Department of Computer Science and Engineering, Methodist College of Engineering and Technology, Hyderabad, India*

Abstract: ***Health insurance fraud has become a serious issue due to the increasing number of digital claims and the involvement of multiple entities such as hospitals, doctors, and patients. Traditional methods often analyze claims individually and fail to capture hidden relationships that may indicate collusion. In this work, a hybrid approach is proposed to improve fraud detection by combining graph-based feature engineering with machine learning techniques. A network is constructed to represent interactions between patients, physicians, and hospitals, from which relational features such as connectivity and visit frequency are derived. In addition, an Isolation Forest model is used to identify unusual financial patterns in claims. These features are then used to train an XGBoost classifier to distinguish between genuine and fraudulent claims. The system also incorporates SHAP-based explanations to provide transparency in predictions. The results indicate that incorporating relational and anomaly-based features improves detection performance compared to traditional models. The proposed approach offers a practical and interpretable solution for identifying both individual and coordinated fraud in healthcare insurance systems.***
Index Terms: Health Insurance Fraud, Graph-Enhanced XGBoost, Collusive Fraud Detection, Explainable AI (XAI), Class Imbalance.

I. INTRODUCTION

Health insurance plays an important role in providing financial support during medical emergencies. However, with the rapid growth of digital claim processing systems, the number of fraudulent claims has also increased significantly. Fraudulent activities such as false billing, upcoding, and coordinated claims between hospitals and patients lead to major financial losses for insurance companies and affect the overall efficiency of the healthcare system.

Most existing fraud detection methods focus on analyzing individual claims using rule-based systems or basic machine learning models. While these approaches can identify simple anomalies, they often fail to detect more complex patterns, especially when multiple entities are involved. In real-world scenarios, fraud is often carried out through collaboration between doctors, hospitals, and patients, making it difficult to detect using traditional techniques. To address this issue, this project proposes a hybrid fraud detection approach that combines graph-based feature engineering with machine learning. By representing interactions between patients, physicians, and hospitals as a network, it becomes possible to capture hidden relationships and identify suspicious patterns. In addition, anomaly detection techniques are used to detect unusual financial behavior in claims.

The extracted features are then used to train an XGBoost model for accurate classification of fraudulent and genuine claims. Furthermore, explainable AI techniques are incorporated to provide transparency in the model's decisions. The overall system is designed to support real-time claim analysis through an interactive interface, making it suitable for practical implementation.

This approach aims to improve fraud detection accuracy while maintaining interpretability, thereby assisting investigators in identifying both individual and collusive fraud cases effectively.

A. Background

Health insurance fraud has become a significant concern in modern healthcare systems due to the increasing volume of digital claims and complex interactions between multiple entities. Fraudulent activities such as false claims, overbilling, and coordinated collusion between healthcare providers and patients lead to substantial financial losses and reduce the efficiency of insurance systems. In many cases, fraudulent behavior is not limited to individual claims but involves networks of entities working together, making detection more challenging. Traditional fraud detection methods primarily focus on rule-based systems or individual claim analysis, which often fail to capture hidden relationships and coordinated patterns.

With the advancement of data-driven techniques, there is a growing need for intelligent systems that can analyze both financial irregularities and relational interactions. Graph-based analysis and machine learning approaches offer a promising direction for identifying such complex fraud patterns in a scalable and interpretable manner

B. The Need / Motivation

Health insurance systems process a large number of claims on a daily basis, making manual verification time-consuming and inefficient. Traditional fraud detection methods rely heavily on predefined rules and human investigation, which are not sufficient to handle the growing complexity of fraudulent activities. In many cases, fraud is carried out through coordinated actions between multiple entities such as patients, doctors, and hospitals, making it difficult to detect using conventional approaches.

Additionally, existing systems often analyze claims individually and fail to consider relationships between different entities. This limitation allows fraudulent patterns to remain hidden within the network of interactions. There is a strong need for an intelligent and automated system that can identify both financial irregularities and relational patterns in real time.

The motivation behind this work is to develop a reliable and efficient fraud detection framework that combines machine learning, graph-based analysis, and anomaly detection. Such a system can assist insurance companies in reducing financial losses and improving the accuracy of fraud identification while maintaining transparency in decision-making.

C. Existing Work

In recent years, several approaches have been proposed for detecting fraud in health insurance systems. Traditional methods are mainly based on rule-based systems and statistical analysis, where predefined conditions are used to identify suspicious claims. While these methods are simple to implement, they are not flexible enough to adapt to new and evolving fraud patterns.

Machine learning techniques such as logistic regression, decision trees, and random forests have been widely used to improve fraud detection. These models learn patterns from historical data and provide better performance compared to rule-based systems. However, most of these approaches treat each claim independently and fail to capture relationships between entities involved in the claim process.

More recent studies have explored graph-based approaches, where interactions between patients, doctors, and hospitals are modeled as networks. These methods are effective in identifying collusive fraud by analyzing connectivity and interaction patterns. In addition, anomaly detection techniques such as Isolation Forest have been used to identify unusual financial behavior in claims.

Despite these advancements, many existing systems either lack interpretability or require high computational resources. This creates a need for a balanced approach that is both efficient and explainable while being capable of detecting complex fraud patterns.

D. Proposed Approach

The proposed system is designed to detect fraudulent health insurance claims by combining relational analysis, anomaly detection, and machine learning techniques. The overall approach focuses on identifying both individual fraudulent claims and coordinated fraud involving multiple entities.

Initially, claim data is processed to extract relevant features such as patient details, hospital information, diagnosis, procedures, and financial attributes. Time-based features, including length of stay, are also derived to capture behavioral patterns in claims.

To capture relationships between entities, a graph-based model is constructed where patients, physicians, and hospitals are represented as nodes, and their interactions are represented as edges. From this network, structural features such as centrality and interaction frequency are extracted. These features help in identifying unusual connectivity patterns that may indicate collusion.

In parallel, an Isolation Forest model is applied to detect anomalies in financial data, such as unusually high claim amounts or abnormal billing patterns. This step helps in identifying claims that deviate from normal behavior.

The extracted features are then combined and used to train an XGBoost classifier for fraud detection. The model learns complex patterns from both relational and financial data to accurately classify claims as genuine or fraudulent. Additionally, explainable AI techniques are incorporated to provide insights into the model's decisions, improving transparency and trust in the system.

II. RESEARCH METHODOLOGY

The methodology adopted in this study focuses on developing a practical and efficient system for detecting fraudulent health insurance claims. The overall approach combines data preparation, feature engineering, anomaly detection, and machine learning to identify both individual and collusive fraud patterns.

Initially, a dataset of health insurance claims is prepared to simulate real-world scenarios. The dataset includes details related to patients, hospitals, physicians, diagnosis, procedures, and financial expenses. To make the data more realistic, both genuine and fraudulent cases are incorporated, with fraud patterns such as unusually high claim amounts, excessive billing, and repeated interactions between specific entities.

In the next step, relevant features are extracted from the dataset. These include financial attributes like total claim amount and number of bills, as well as time-based features such as length of hospital stay. In addition to this, relationships between patients, doctors, and hospitals are modeled using a graph structure, which helps in identifying unusual interaction patterns that may indicate collusion.

To further strengthen the detection process, an anomaly detection technique is applied using the Isolation Forest algorithm. This helps in identifying claims that deviate significantly from normal behavior. The anomaly scores generated are then used as additional inputs for the classification model.

Finally, an XGBoost classifier is trained using the combined set of features. The model is evaluated using standard performance metrics such as precision, recall, F1-score, and ROC-AUC. This structured methodology ensures that the system can effectively detect fraudulent claims while maintaining reliability and interpretability.

A. Data Description

The data used in this study is synthetically generated to simulate real-world health insurance claim scenarios. The dataset contains detailed information about patients, hospitals, physicians, diagnosis, procedures, and financial transactions associated with each claim.

Each claim record includes attributes such as claim reference number, policy number, hospital name, attending physician, diagnosis, procedure performed, and various expense components including pre-hospitalization, inpatient, and post-hospitalization costs. Time-related attributes such as admission and discharge dates are also included to capture behavioral patterns.

The dataset is designed to include both normal and fraudulent claim patterns. Fraudulent instances are introduced through scenarios such as unusually high claim amounts, excessive number of bills, and coordinated interactions between specific hospitals and physicians. This helps in evaluating the effectiveness of the proposed fraud detection system under realistic conditions.

The generated dataset provides sufficient variability and complexity, making it suitable for training and testing machine learning models for fraud detection.

B. Feature Engineering and Model Framework

In order to effectively detect fraudulent claims, multiple features are engineered from the raw dataset to capture both financial behavior and relational patterns. These features play a crucial role in improving the performance of the proposed fraud detection system. Initially, time-based features such as length of stay are derived from admission and discharge dates. Financial features including total claim amount, number of bills submitted, and expense breakdown (pre-hospitalization, inpatient, and post-hospitalization) are also considered. Additionally, derived features such as cost per day are calculated to identify abnormal billing patterns. To capture relationships between entities, a graph-based representation is constructed using patients, physicians, and hospitals as nodes. Interactions between these entities are represented as edges. From this graph, structural features such as degree centrality and interaction frequency are extracted. These features help in identifying unusual connectivity patterns that may indicate collusion. Furthermore, an Isolation Forest model is used to generate anomaly scores based on financial attributes. This helps in identifying claims that significantly deviate from normal patterns. The final feature set, combining financial, behavioral, relational, and anomaly-based features, is used to train an XGBoost classifier. This integrated approach enables the system to effectively distinguish between genuine and fraudulent claims while maintaining computational efficiency.

C. Overview of Machine Learning for Fraud Detection

Machine learning plays an important role in identifying patterns and anomalies in large datasets. In the context of health insurance fraud detection, machine learning models are used to analyze claim data and distinguish between genuine and fraudulent cases based on learned patterns.

Unlike traditional rule-based systems, which rely on predefined conditions, machine learning models can adapt to new and evolving fraud behaviors by learning from historical data. These models consider multiple factors such as claim amount, number of bills, duration of hospital stay, and relationships between different entities involved in the claim.

Tree-based models such as Random Forest and XGBoost are particularly effective for structured data, as they can handle non-linear relationships and interactions between features. In addition, anomaly detection techniques such as Isolation Forest are used to identify unusual claim patterns that deviate from normal behavior.

By combining these techniques, the system can effectively detect both known fraud patterns and previously unseen anomalies, making it a reliable approach for real-world fraud detection applications.

D. Basics of Fraud Detection in Health Insurance

Fraud in health insurance refers to the act of making false or misleading claims to obtain financial benefits. It can occur in various forms, such as overbilling, unnecessary medical procedures, duplicate claims, or coordinated activities between patients and healthcare providers. These fraudulent practices result in significant financial losses and reduce the efficiency of insurance systems.

Detecting fraud is challenging because fraudulent claims often resemble genuine ones and may involve multiple entities working together. Traditional methods rely on manual verification and rule-based systems, which are time-consuming and not effective for identifying complex patterns.

With the advancement of data-driven techniques, fraud detection has increasingly shifted towards machine learning approaches. These methods analyze patterns in claim data and identify unusual behavior based on multiple attributes such as claim amount, number of bills, and duration of hospital stay. In addition, relational analysis using graph-based techniques helps in detecting collusive fraud by examining interactions between patients, doctors, and hospitals.

By combining statistical analysis, anomaly detection, and machine learning models, modern systems can improve the accuracy and efficiency of fraud detection in healthcare insurance.

III. LITERATURE REVIEW

s.no	Paper Title	Authors	Year	Methodology	Dataset Used	Models Used	Research Gaps
1	Robust interpretable ensemble for healthcare insurance fraud	Wang et al.	2025	Ensemble boosting + feature selection + SHAP/LIME	US claims (10K synthetic/real)	XGBoost, LightGBM, CatBoost, RF	No graph relations; no drift adaptation; limited deployment
2	Explainable unsupervised anomaly detection for insurance data	De Meulemeester et al.	2025	Unsupervised anomaly ranking + SHAP workflow	Belgian provider claims (~50K)	VAE, OCSVM, LOF, KNN, COPOD	No per-claim scoring; no graphs; no real-time deploy
3	Healthcare fraud detection using adaptive DL	Matloob et al.	2025	Association rules + anomaly transformer cascade	Hospital logs (2013-2019)	Rule engine + Anomaly Transformer	High compute; low explainability; not lightweight
4	Automating claims with supervised/unsupervised AI	Hassan & Alam	2024	LR + autoencoder fusion for classification	Anonymized claims (1,200)	LR + Deep Autoencoder	Small dataset; no graphs; limited interpretability
5	Healthcare Insurance Fraud Detection Using ML	Chitteti et al.	2025	Supervised classification + anomaly isolation	Indian claims (~5K anonymized)	CatBoost + Isolation Forest	No SHAP; no collusion graphs; no drift handling

6	AI techniques for healthcare fraud detection (survey)	Razzaq et al.	2025	PRISMA review of ML/DL trends	Global datasets (Medicare/Kaggle)	Survey: DT/RF/SVM, clustering, CNN/LSTM, XGBoost	Imbalance/explainability gaps; no deploy frameworks
7	ML approaches for health insurance fraud detection	Sharma	2024	ML/XAI review + benchmarks	Mixed US/India claims (~20K)	LR, DT, RF, ensembles + SHAP/LIME	No cross-dataset generalization; no end-to-end graphs
8	Explainable anomaly workflow for insurance	De Meulemeester et al.	2025	Anomaly ranking + SHAP visuals	Belgian claims/providers (~100K)	VAE + LOF/KNN ensemble	No graph collusion; no adaptive drift
9	Multi-channel heterogeneous graph for health fraud	Li et al.	2024	Heterogeneous GNN + channel fusion	Chinese claims/providers (~200K)	MHGSL (Heterogeneous GNN)	High complexity/compute; no interpretability
10	Fraud detection using SMOTE-Boruta and ML	Nabrawi et al.	2023	SMOTE + Boruta + supervised classification	Saudi claims (~15K anonymized)	RF, LR, ANN	Tabular only; no graphs/drift
11	Autoencoder-based fraud detection for claims	Alghamdi et al.	2022	Autoencoder reconstruction for anomalies	US synthetic claims (~50K)	Deep Variational Autoencoder	No interpretability; unsupervised only; no workflow
12	CNN-LSTM + SHA-256 for secure fraud detection	Kumar et al.	2023	CNN-LSTM sequences + encryption	Hospital logs (~8K de-identified)	CNN-LSTM + SHA-256	No explainability; no graphs; complex deploy
s.no	Paper Title	Authors	Year	Methodology	Dataset Used	Models Used	Research Gaps
13	Next-Gen ML in Healthcare Fraud Detection	Alghamdi et al.	2025	Hybrid DL survey + XAI benchmarks	Mixed (Medicare/Kaggle/MIMIC-III)	CNN/LSTM + SHAP	No India adaptations; limited graph collusions
14	Predictive Accuracy of ML in Indian Health Fraud	Aditi Sharma	2024	Desk-review + supervised metric evals	Indian claim studies (~10K aggregated)	XGBoost/RF + SMOTE	No relational graphs; no drift adaptation
15	Lightweight Graph-Enhanced XGBoost for Collusive Fraud in Indian Health Insurance (Our Proposed Work)	Proposed Work	2025	Graph features + anomaly scores + XGBoost + SHAP + Streamlit	Indian synthesized data (~600K)	XGBoost + RF + Isolation Forest	Fills all: Lightweight graphs for relations, imbalance ablations, SHAP explainability.

IV. PROPOSED SYSTEM

The proposed system aims to detect fraudulent health insurance claims by integrating machine learning, graph-based analysis, and anomaly detection techniques. The system is designed to identify both individual fraudulent claims and coordinated collusive fraud involving multiple entities such as patients, physicians, and hospitals.

The architecture consists of multiple stages, beginning with the input of structured claim data through a standardized e-claim form. The input data includes patient details, hospital information, diagnosis, procedures, and financial attributes. This data is preprocessed to ensure consistency, including encoding categorical variables and handling missing values.

In the next stage, feature engineering is performed to extract meaningful insights from the data. This includes calculating time-based features such as length of stay and financial indicators such as total claim amount and cost per day. Additionally, a graph-based representation is constructed to model relationships between patients, doctors, and hospitals. From this graph, relational features such as centrality and interaction frequency are derived to identify suspicious connectivity patterns.

To enhance detection capability, an anomaly detection module based on the Isolation Forest algorithm is applied to identify unusual financial behavior. These anomaly scores are combined with other features and passed to an XGBoost classifier, which predicts whether a claim is fraudulent or genuine.

To ensure transparency, the system incorporates SHAP-based explainability, which highlights the key factors influencing each prediction. The final output includes fraud classification, risk score, and explanatory insights, all presented through an interactive dashboard that enables real-time analysis and decision-making.

V. METHODOLOGY

The methodology of the proposed system is designed to effectively detect fraudulent health insurance claims by combining data preprocessing, feature engineering, anomaly detection, and machine learning techniques. The approach focuses on identifying both individual fraud and collusive fraud patterns present in claim data.

Initially, a structured dataset of health insurance claims is prepared, containing information related to patients, hospitals, physicians, diagnosis, procedures, and financial transactions. The data undergoes preprocessing, where missing values are handled, categorical variables are encoded, and numerical features are scaled where necessary to ensure consistency and improve model performance.

In the next stage, feature engineering is performed to extract meaningful attributes from the raw data. Time-based features such as length of stay are derived from admission and discharge dates. Financial indicators including total claim amount and cost per day are calculated to identify abnormal billing patterns. In addition, a graph-based representation is constructed to model relationships between patients, physicians, and hospitals. From this graph, relational features such as degree centrality and interaction frequency are extracted to detect potential collusion.

To further enhance the detection capability, an anomaly detection technique is applied using the Isolation Forest algorithm. This model identifies claims that significantly deviate from normal behavior based on financial and temporal features. The anomaly scores generated are incorporated as additional inputs for the classification model.

Finally, an XGBoost classifier is trained using the combined set of features. To address the issue of class imbalance, SMOTE is applied to the training data, ensuring that the model learns patterns from both genuine and fraudulent claims effectively. The performance of the model is evaluated using metrics such as precision, recall, F1-score, and ROC-AUC.

This multi-stage methodology enables the system to capture complex fraud patterns while maintaining efficiency and interpretability, making it suitable for real-world deployment.3.4.3 Comparison of the Models

VI. IMPLEMENTATION

The implementation of the proposed fraud detection system is carried out using Python due to its flexibility and strong ecosystem of data processing and machine learning libraries. The system integrates multiple components, including data preprocessing, feature engineering, anomaly detection, and classification, to create an end-to-end fraud detection pipeline.

The dataset is first processed using libraries such as Pandas and NumPy for efficient handling of large-scale structured data. Categorical variables, including diagnosis, procedure, and network status, are encoded using label encoding techniques, while numerical features are standardized to improve model performance. The dataset is then divided into training and testing sets to evaluate the effectiveness of the model. To address class imbalance, SMOTE (Synthetic Minority Oversampling Technique) is applied to the training data, ensuring that the model learns patterns from both genuine and fraudulent claims. For anomaly detection, the Isolation Forest algorithm is implemented to identify unusual financial patterns based on features such as claim amount, number of bills, and length of stay. The anomaly scores generated are incorporated into the feature set.

The core classification model is built using XGBoost, which is well-suited for structured data and capable of handling complex relationships between features. The model is trained using the engineered dataset and optimized for accurate fraud detection. Evaluation metrics such as precision, recall, F1-score, and ROC-AUC are used to assess performance.

To enhance interpretability, SHAP (SHapley Additive exPlanations) is integrated into the system to explain individual predictions by highlighting the contribution of each feature. The entire system is deployed using a Streamlit interface, allowing users to upload claim data, perform real-time analysis, and visualize results in an interactive manner.

This implementation ensures that the system is scalable, efficient, and capable of providing accurate and explainable fraud detection in practical scenarios

VII. CHALLENGES

A. *Dataset Imbalance*

One of the major challenges in fraud detection is the imbalance in the dataset, where fraudulent claims represent only a small percentage of the total data. This imbalance can lead to biased model learning, where the model tends to favor genuine claims. Techniques such as SMOTE are used to address this issue, but maintaining a balance between precision and recall remains a challenge.

B. *Detection of Collusive Fraud*

Fraud is often not limited to individual claims but involves coordinated activities between patients, physicians, and hospitals. Detecting such collusive behavior is difficult because individual claims may appear normal, and only relational patterns reveal suspicious activity.

C. *Data Quality and Consistency*

Real-world insurance data may contain missing, inconsistent, or noisy information. Ensuring data quality through preprocessing and feature engineering is essential but challenging, especially when dealing with large-scale datasets.

D. *Evolving Fraud Patterns*

Fraudsters continuously adapt their strategies to bypass detection systems. As a result, models trained on historical data may become less effective over time, requiring periodic retraining and updates.

E. *False Positives and False Negatives*

Balancing precision and recall is a key challenge. A high number of false positives can increase investigation costs, while false negatives may allow fraudulent claims to go undetected.

F. *Computational Complexity*

Handling large datasets and graph-based computations can increase processing time and resource requirements. Ensuring scalability while maintaining performance is an important consideration.

G. *Explainability and Trust*

Although machine learning models can achieve high accuracy, their decisions must be interpretable. Providing clear explanations for predictions is essential for gaining trust from investigators and stakeholders.

VIII. RESULTS AND DISCUSSION

The proposed fraud detection system was evaluated using multiple machine learning models, including Logistic Regression, Random Forest, and the proposed Graph-Augmented XGBoost model. The evaluation was carried out using standard performance metrics such as precision, recall, F1-score, and ROC-AUC to assess the effectiveness of each model.The experimental results indicate that Logistic Regression achieves a high recall but suffers from very low precision, which suggests that it incorrectly classifies a large number of genuine claims as fraudulent. Random Forest improves overall performance by increasing precision while maintaining high recall, providing a better balance compared to Logistic Regression.

The proposed Graph-Augmented XGBoost model achieves the best performance in terms of F1-score, indicating an improved balance between precision and recall. Although the recall is slightly lower than other models, it significantly reduces false positives,

making it more practical for real-world deployment. The ROC-AUC values across all models are high, demonstrating strong discriminative capability.The improved performance of the proposed model can be attributed to the integration of graph-based relational features and anomaly detection scores. These additional features enable the model to capture complex fraud patterns, including collusive behavior between entities, which cannot be identified using traditional tabular approaches alone.Overall, the results demonstrate that the proposed system provides a reliable and efficient solution for detecting health insurance fraud, with improved accuracy and better interpretability compared to conventional methods.

```
Confusion Matrix:
[[112765   4780]
 [   372   2083]]

Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.96      0.98    117545
           1       0.30      0.85      0.45      2455

    accuracy                           0.96    120000
   macro avg       0.65      0.90      0.71    120000
weighted avg       0.98      0.96      0.97    120000

ROC-AUC Score: 0.9782
```

IX. FUTURE SCOPE

Integration of Graph Neural Networks (GNNs) While graph features are used in this project, future work can explore GNN-based models such as GraphSAGE or GCN to automatically learn deeper relational patterns.

Real-Time Fraud Monitoring The system can be extended into a real-time detection platform that flags suspicious behaviour as soon as a claim is submitted.

Inclusion of Additional Data Sources Future versions can incorporate:

1) pharmacy records
2) medical referral sequences
3) geographic information
4) social network data

These can help reveal even more complex fraud structures.

Dashboard for Investigators A user-friendly dashboard with interactive graph visualizations and SHAP explanations can help investigators quickly review flagged claims.

Broader Application The same approach can be adapted for:

- banking fraud
- e-commerce fraud
- telecom fraud
- cybersecurity attacks

Collaboration with Insurance Companies Future work can involve deploying the model with actual insurance firms to refine performance on real-world data.

X. CONCLUSION

This work examined the limitations of existing techniques for detecting health-insurance fraud, particularly in cases involving coordinated and relational behaviours that are not visible at the level of individual claims. Traditional rule-based systems and standalone supervised models struggle with severe class imbalance, evolving fraud patterns and the absence of interpretability required for investigative decision-making. To address these gaps, we proposed a lightweight graph-augmented XGBoost framework that combines tabular claim features with graph-derived relational indicators and anomaly-detection signals. The integration of SHAP explanations ensures transparency, while adaptive retraining supports long-term resilience against drift. Experimental results demonstrate improved fraud-detection performance over baseline models without demanding high computational resources.

Overall, the framework provides a practical, explainable and scalable approach to fraud detection and represents a step toward operational systems that can analyse claims collectively rather than in isolation. Future extensions may include federated collaboration across insurers and enhanced support for unstructured clinical information.

XI. ACKNOWLEDGMENT

We sincerely thank Dr. Prabhu G. Benakop , Principal, and Dr. Lavanya Pamulaparty, Head of the Department of Computer Engineering, Methodist College of Engineering and Technology, for their constant support and for providing the facilities needed to complete this project.

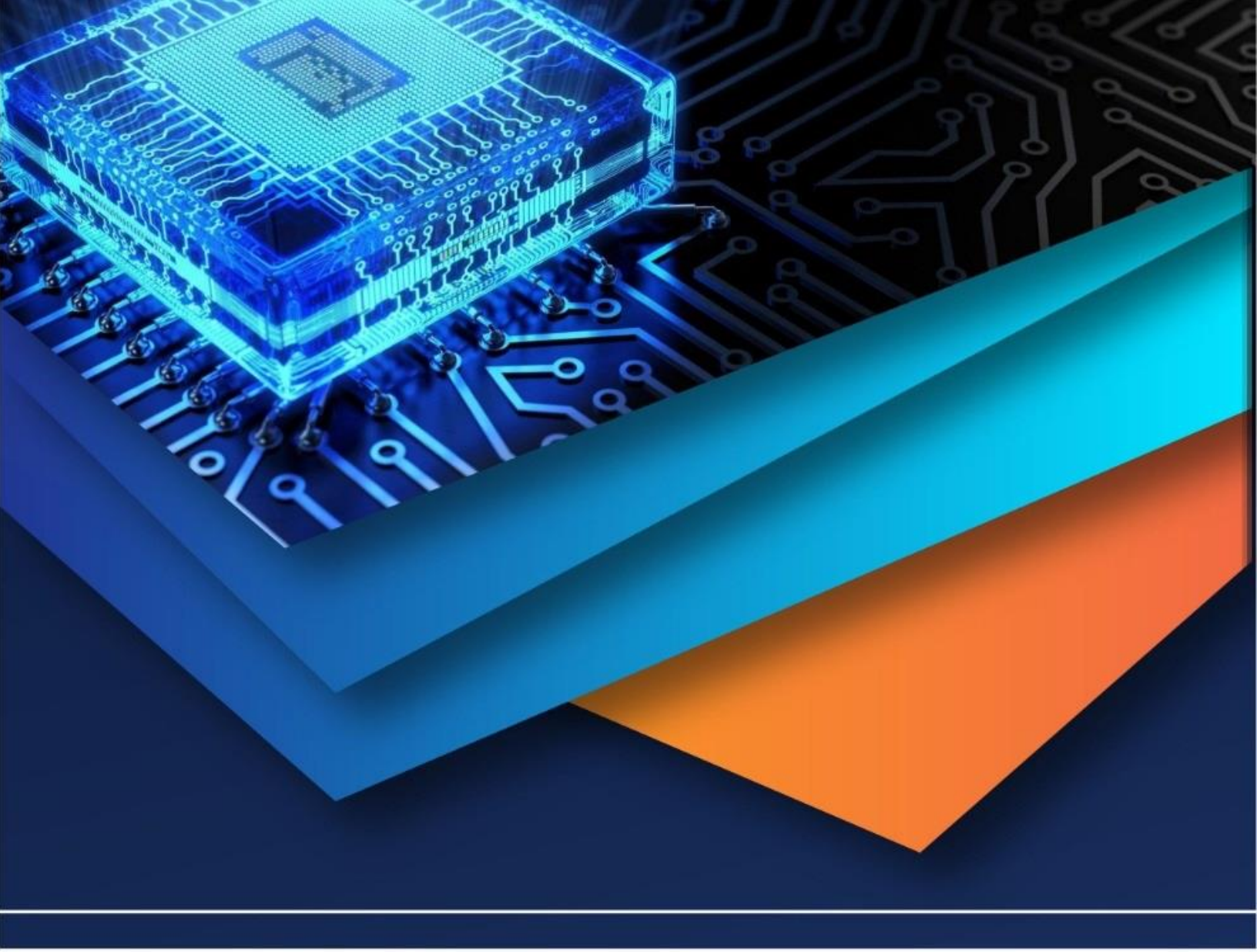
We are especially grateful to our project guide, Mrs. M. Aruna, for her guidance, patience, and encouragement throughout this work.

We also thank our project coordinator, Dr. Shahana Tanveer, for her support and valuable suggestions.

Finally, we appreciate all the faculty and staff members for their help and cooperation during the course of this project.

REFERENCES

[1] J. Wang et al., "A robust and interpretable ensemble ML model for predicting healthcare insurance fraud," Expert Syst. Appl., vol. 250, p. 124567, Aug. 2025.
[2] J. De Meulemeester et al., "Explainable unsupervised anomaly detection for healthcare insurance data," BMC Med. Inform. Decis. Mak., vol. 25, no. 1, p. 823, 2025.
[3] A. Matloob et al., "Healthcare fraud detection using adaptive and deep learning," Neural Comput. Appl., vol. 37, no. 12, pp. 9698-9710, 2025.
[4] J. Hassan and M. Alam, "Automating healthcare claims with supervised and unsupervised AI," Appl. Sci., vol. 14, no. 5, p. 2100, Mar. 2024.
[5] S. Chitteti, P. P. Shenoy, and P. K. Vidhate, "Healthcare insurance fraud detection using ML," SN Comput. Sci., vol. 6, no. 3, p. 456, May 2025.
[6] K. Razzaq, M. Shah, and A. Alghamdi, "AI techniques for healthcare fraud detection: A survey," Information, vol. 16, no. 9, p. 730, Sep. 2025.
[7] A. Sharma, "ML approaches for health insurance fraud detection," Int. J. Innov. Sci. Res. Technol., vol. 9, no. 6, pp. 1234-1245, Jun. 2024.
[8] J. De Meulemeester et al., "Explainable anomaly workflow for healthcare insurance," BMC Med. Inform. Decis. Mak., vol. 25, no. 1, p. 824, 2025.
[9] H. Li et al., "Health insurance fraud detection based on multi-channel heterogeneous graph structured learning," Appl. Soft Comput., vol. 150, p. 111616, Oct. 2024.
[10] M. Nabrawi et al., "Fraud detection in healthcare insurance claims using machine learning," Risks, vol. 11, no. 9, p. 160, Sep. 2023.
[11] A. Alghamdi et al., "Autoencoder-based fraud detection for insurance claims," IEEE Access, vol. 10, pp. 56789-56800, 2022.
[12] R. Kumar et al., "CNN-LSTM + modified SHA-256 for secure healthcare fraud detection," Wireless Pers. Commun., vol. 128, no. 2, pp. 234-250, Jan. 2023.
[13] A. Alghamdi et al., "Next-generation machine learning in healthcare fraud detection," Information, vol. 16, no. 9, p. 730, Sep. 2025.
[14] A. Sharma, "Predictive accuracy of ML models in fraud detection for health insurance in India," Amer. J. Soc. Sci. Admin. Stud., vol. 3, no. 2, p. 2253, 2024.
[15] Insurance Regulatory and Development Authority of India (IRDAI), "Annual Report on Health Insurance Fraud and Abuse," IRDAI, Hyderabad, India, 2025.
[16] Federation of Indian Chambers of Commerce & Industry (FICCI), "Working Paper on Health Insurance Abuse and Fraud Management," FICCI Sub-Group on Health Insurance Fraud, New Delhi, India, 2025.
[17] BCG and Medi Assist, "Rebuilding Trust in India's Health Insurance Ecosystem: Tackling Fraud and Abuse," BCG Report, Mumbai, India, Nov. 2025.
[18] Centers for Medicare & Medicaid Services (CMS), "Improper Payments in Medicare and Medicaid Programs," U.S. Dept. Health Human Services, Washington, DC, USA, 2024.
[19] C. Phua, V. Lee, K. Smith, and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," Artif. Intell. Rev., vol. 33, no. 1, pp. 1-30, 2010 (updated 2025 ed.).
[20] J. N. Trivedi and R. Vagadiya, "Artificial intelligence in the detection and prevention of insurance fraud in India," Int. Educ. J. Soc. Sci. Educ., vol. 1, no. 1, p. 165, 2025.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY