© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717969

A Comprehensive Review and Evaluation Framework for Explainable Healthcare Insurance Fraud Detection Models

SHRAVYA CAROLINE PALLAT[1], RAKSHITHA B S[2]
[1]*MSC-CSIT, Department of Computer Science and IT, Jain (Deemed to be) University Bangalore, India*
[2]*Assistant Professor, Department of Computer Science and IT, Jain (Deemed to be) University Bangalore, India*

Abstract- ***Healthcare insurance fraud poses a significant challenge to healthcare systems worldwide, resulting in substantial financial losses and inefficiencies. Traditional detection methods have proven inadequate in handling large-scale and complex fraud patterns, leading to the adoption of machine learning and deep learning techniques. Existing studies have focused primarily on improving detection accuracy using various models such as random forests, neural networks, and hybrid approaches. While these techniques demonstrate strong predictive performance, they often operate as black-box systems and lack transparency. Additionally, there is limited comparative analysis of explainability techniques across models. This study identifies a critical research gap in the lack of auditor-centric explainability and absence of frameworks that balance accuracy with interpretability. To address this, the paper proposes an evaluation framework that assesses multiple fraud detection models based on both performance metrics and interpretability criteria. The proposed approach involves applying selected models to healthcare datasets, analyzing their predictions using explainability techniques, and comparing them based on usability for real-world auditing. The expected outcomes include identifying models that provide both accurate and interpretable results. This research contributes by providing a structured literature review, identifying key gaps, and proposing a practical evaluation framework to enhance transparency and decision-making in healthcare fraud detection systems.***

Index Terms - Healthcare Fraud Detection, Machine Learning, Explainable AI, Interpretability, Fraud Detection Models, Healthcare Analytics

I. INTRODUCTION

I. Background of the Study

Healthcare insurance plays a critical role in ensuring access to medical services while reducing financial burden on individuals. However, increasing volume of healthcare transactions has also led to a significant rise in fraudulent activities, including false claims, overbilling, and collusion between stakeholders. These fraudulent practices result in substantial financial losses for insurance providers and negatively impact overall efficiency of healthcare systems.

With the digitization of healthcare records and insurance processes, large volumes of data are generated in the form of electronic health records and insurance claims. This has created opportunities to apply advanced analytical techniques such as machine learning and data mining for fraud detection. These approaches enable automated identification of suspicious patterns and improve detection accuracy compared to traditional manual methods.

Despite these advancements, healthcare fraud detection remains a complex challenge due to the dynamic nature of fraudulent behavior, data imbalance, and the need for real-time decision-making. Therefore, developing effective and practical fraud detection solutions is essential for improving transparency, reducing financial losses, and ensuring the sustainability of healthcare systems.

II. Problem Statement

Although numerous machine learning and deep learning techniques have been proposed for healthcare fraud detection, most existing approaches focus primarily on improving predictive accuracy

while overlooking interpretability and usability. Many advanced models, particularly deep learning and graph-based approaches, operate as black-box systems, making it difficult for auditors and decision-makers to understand the reasoning behind fraud predictions.

Additionally, current studies lack systematic comparison of explainability techniques across different models. In particular, limited studies comparatively evaluate SHAP and LIME techniques for healthcare insurance fraud detection models. The absence of auditor-centric frameworks limits the practical adoption of these systems in real-world environments, where transparency and accountability are essential.

Therefore, the key problem addressed in this study is lack of fraud detection frameworks that balance high performance with interpretability and real-world usability, particularly for healthcare auditing purposes.

III. Motivation

The motivation for this study arises from growing need for reliable and transparent fraud detection systems in healthcare. As fraud schemes become more sophisticated, traditional detection methods are no longer sufficient to handle complex patterns and large-scale data.

At the same time, regulatory requirements and industry practices increasingly demand explainable and trustworthy AI systems. Decision-makers and auditors require not only accurate predictions but also clear explanations to justify actions and ensure compliance.

This study is motivated by the need to bridge the gap between high-performance fraud detection models and their practical usability, ensuring that these systems can be effectively applied in real-world healthcare environments

IV. Objectives of the Study

Objective 1: To review existing approaches for healthcare fraud detection, including machine learning, deep learning, and hybrid methods.

Objective 2: To identify research gaps, particularly in terms of interpretability and usability of fraud detection models.

Objective 3: To propose an evaluation framework that compares fraud detection models based on both performance and explainability.

V. Contributions of the Paper

This paper makes the following contributions:

- Provides comprehensive review of existing healthcare fraud detection techniques, covering traditional, machine learning, deep learning, and advanced approaches.
- Identifies critical research gaps, including the lack of auditor-centric explainability and the absence of systematic evaluation of interpretability techniques.
- Proposes a novel evaluation framework that assesses fraud detection models based on both predictive performance and interpretability.

These contributions aim to support the development of more practical and transparent fraud detection systems in healthcare.

II. RELATED WORK AND LITERATURE REVIEW

I. Thematic Classification of Literature

1. Traditional Approaches: Early healthcare fraud detection relied on rule-based systems and manual auditing. These systems lacked scalability and adaptability and struggled to detect complex and evolving fraud patterns. This leads to high false-positive rates. As fraud schemes became more sophisticated, these approaches were slowly replaced by data-driven techniques [18], [22].

2. Machine Learning Approaches: Machine learning techniques like decision trees, logistic regression, and random forests have been widely adopted for fraud detection tasks. These models learn patterns from historical data and classify claims as fraudulent or legitimate. Several studies show high accuracy using supervised learning models, especially combined with feature selection and class balancing techniques. Traditional interpretable machine learning models

remain practically important due to their compatibility with post-hoc explainability techniques such as SHAP and LIME. [1], [5], [3].

Despite the effectiveness, models rely heavily on labelled datasets and often fail to capture complex relationships between entities such as patients and providers. Most traditional ML models lack interpretability, making it difficult for auditors to understand the reasoning behind predictions [2], [4].

3. Deep Learning Approaches: Deep learning techniques, such as neural networks and attention-based models, have been researched to improve fraud detection performance. These models can capture complex, non-linear relationships in large datasets and have demonstrated improved detection accuracy [8], [21]. However, deep learning models are computationally intensive and require large datasets for training. More importantly, they function as black-box systems, providing limited transparency in decision-making, which restricts their adoption in real-world healthcare auditing environments [7].

4. Data Mining and Unsupervised Approaches: Unsupervised learning and data mining techniques, such as clustering and association rule mining, are commonly used when labelled data is unavailable. These approaches help identify anomalies and hidden patterns in claims data [19], [18]. Although effective in detecting unusual behavior, these methods often struggle to distinguish between legitimate anomalies and fraudulent activities. Furthermore, the results are not easily interpretable, limiting their practical application in decision-making scenarios [25].

5. Hybrid and Ensemble Models: Hybrid and ensemble models combine multiple algorithms to improve fraud detection performance. Techniques such as stacking and boosting leverage the strengths of individual models to achieve better accuracy and robustness [9], [10]. While these approaches enhance predictive performance, they increase model complexity and reduce interpretability. This makes it difficult for stakeholders to understand the reasoning behind fraud detection outcomes.

6. Blockchain and Security-Based Approaches: Recent studies have explored the integration of blockchain technology with machine learning to improve data security and transparency in healthcare fraud detection systems. Blockchain ensures immutability and secure data sharing, while machine learning enables intelligent fraud detection [12], [14]. However, these systems primarily focus on security and data integrity rather than explainability. Although they enhance transparency at the data level, they do not provide clear explanations for fraud detection decisions [11], [13].

7. Graph-Based and Advanced Approaches: Graph-based approaches, particularly Graph Neural Networks (GNNs), have emerged as a promising solution for detecting complex fraud patterns involving multiple entities. These models capture relationships between patients, providers, and claims, enabling detection of collusive fraud networks [17]. Despite their effectiveness, graph-based models are computationally expensive and lack interpretability. Understanding how decisions are made within these models remains a significant challenge.

Although advanced approaches demonstrate strong performance, this study primarily focuses on interpretable machine learning models due to their practical suitability for explainability analysis and auditing.

II. Comparative Analysis of Existing Methods

TABLE 1. COMPARATIVE ANALYSIS OF EXISTING METHODS

Author	Year	Method	Dataset	Performance	Limitations
Nabrawi et al. [1]	2023	RF, LR, ANN	Healthcare claims	~98% accuracy	Requires labelled data, low interpretability
Hamid et al. [19]	2024	Rule mining + anomaly detection	CMS dataset	Moderate	Low explainability

Hasan [17]	2022	GNN	Relational healthcare data	High accuracy	Complex, low transparency
Amponsah et al. [14]	2022	DT + Blockchain	Claims dataset	~97%	Focus on security
Agarwal [25]	2023	K-means clustering	Claims data	Improved detection	Weak anomaly distinction
Dey et al. [2]	2025	ML + anomaly detection	EHR + claims	High accuracy	Privacy + interpretability issues
Debener [3]	2023	Hybrid ML	Insurance dataset	Balanced	Scalability issues

III. CRITICAL REVIEW

The reviewed literature demonstrates that machine learning and deep learning techniques significantly improve fraud detection accuracy. Ensemble and hybrid approach further enhance performance by combining multiple models [9], [10].

However, several limitations persist. Most models lack interpretability, making them unsuitable for real-world auditing applications. Deep learning and graph-based approaches, although highly accurate, operate as black-box systems, limiting trust and transparency [7], [17].

Scalability remains another concern, as many models are computationally expensive and not designed for large-scale deployment. Additionally, most studies rely on imbalanced or limited datasets, which affects generalization across different healthcare systems [5], [19].

Furthermore, blockchain-based systems improve security but do not address explainability, while unsupervised approaches often produce ambiguous results that are difficult to validate [12], [25].

IV. IDENTIFIED RESEARCH GAPS

Based on the literature, the following research gaps are identified:

- Lack of systematic comparison of explainability techniques across models
- Absence of auditor-centric fraud detection frameworks
- Limited focus on the trade-off between accuracy and interpretability
- Limited comparative analysis of SHAP and LIME in healthcare fraud detection contexts.

V. PROPOSED METHODOLOGY

I. System Overview

This research proposes an evaluation framework for healthcare fraud detection models, with a focus on interpretability and usability for auditing purposes. Unlike traditional approaches that emphasize only predictive accuracy, the proposed framework evaluates models based on their ability to provide meaningful and understandable explanations for fraud detection decisions.

The system is designed to analyze multiple fraud detection techniques, including selected machine learning models including Logistic Regression, Random Forest, SVM, and XGBoost, and assess their performance using both quantitative metrics (accuracy, precision, recall) and qualitative interpretability measures.

The framework consists of the following components:

- Data Input Layer: Healthcare claims dataset containing structured features such as patient details, billing information, and provider data.
- Model Layer: Pre-existing fraud detection models (e.g., Random Forest, Logistic Regression, Neural Networks).
- Explainability Layer: Application of interpretability techniques such as SHAP (SHapley Additive Explanations), and LIME (Local Interpretable Model-Agnostic Explanations).

- Evaluation Layer: Comparative analysis of models based on performance and interpretability.
- Output Layer: Auditor-friendly insights, including explanations for fraud predictions.

This architecture enables a structured comparison of models while emphasizing their practical usability in real-world fraud detection scenarios.

II. Workflow Diagram

The proposed framework follows a systematic workflow that transforms raw healthcare data into interpretable fraud detection insights.

Step 1: Data Input

Healthcare claims data is collected from publicly available datasets or benchmark datasets. The data includes attributes such as patient demographics, treatment details, billing codes, and claim history.

Step 2: Data Preprocessing

The dataset is cleaned and prepared for analysis. This includes:

- Handling missing values
- Encoding categorical variables
- Normalizing numerical features
- Addressing class imbalance using SMOTE

Step 3: Model Selection and Application

Multiple fraud detection models are applied to the dataset. Selected machine learning models including Logistic Regression, Random Forest, SVM, and XGBoost.

Step 4: Explainability Analysis

Explainability techniques are applied to each model to understand prediction behavior. These may include:

- Feature importance analysis
- SHAP (SHapley Additive Explanations)
- LIME-based local explanations)

Step 5: Evaluation

Models are evaluated based on:

- Performance metrics: Accuracy, precision, recall, F1-score
- Interpretability metrics: Clarity of explanations, feature relevance, and usability

Step 6: Output Generation

The final output includes:

- Fraud prediction results
- Explanation of why a claim is classified as fraudulent
- Comparative insights across different models

This workflow ensures that the system not only detects fraud but also provides meaningful insights for decision-making.

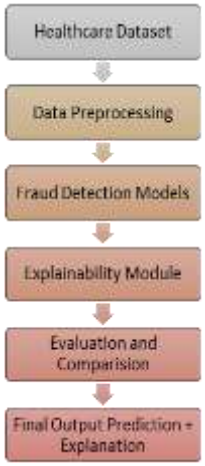

FIGURE 1. WORKFLOW DIAGRAM.

III. Dataset Description

This study proposes to utilize the Synthetic Healthcare Fraud Dataset obtained from Kaggle, which contains approximately 10,000 healthcare insurance claim records. The dataset includes both legitimate and fraudulent claims and was selected due to its suitability for healthcare fraud detection experimentation and explainability analysis.

The dataset contains a combination of numerical and categorical features related to healthcare insurance claims, including billed amount, diagnosis codes, procedure codes, claim status, provider information, and service-related attributes.

Exploratory analysis revealed significant class imbalance, with fraudulent claims representing a minority class compared to legitimate claims. To address this issue, Synthetic Minority Oversampling Technique (SMOTE) needs to be applied during preprocessing to improve fraud detection performance.

The dataset must be preprocessed through:

- handling missing values
- feature encoding
- train-test splitting
- feature scaling (for selected models)
- class balancing using SMOTE

This preprocessing pipeline will ensure compatibility with machine learning models and explainability techniques such as SHAP and LIME.

VI. EXPECTED RESULTS AND DISCUSSIONS

I. Expected Outcomes

The proposed evaluation framework is expected to provide a comprehensive comparison of fraud detection models based on both performance and interpretability.

Firstly, traditional machine learning models such as Random Forest and Logistic Regression are expected to achieve high accuracy and stable performance, especially on structured healthcare datasets. These models are also anticipated to provide relatively better interpretability compared to complex models.

Secondly, deep learning and advanced models are expected to demonstrate higher capability in detecting complex fraud patterns, particularly in large and high-dimensional datasets. However, their interpretability is expected to be limited due to their black-box nature.

The framework is also expected to highlight the trade-off between accuracy and interpretability, showing that models with higher predictive performance may not always provide meaningful explanations.

In terms of robustness, the framework is expected to evaluate how different models perform under varying data conditions, such as class imbalance and noisy data. Traditional models are expected to be more stable, while advanced models may be more sensitive to data variations.

Scalability is also considered, where simpler models are expected to perform efficiently on large datasets, while deep learning and graph-based models may require higher computational resources.

II. Comparative Evaluation Plan

The evaluation will be conducted by applying multiple fraud detection models to the selected dataset and comparing their performance using both quantitative and qualitative criteria.

Performance Metrics

- Accuracy
- Precision
- Recall
- F1-score

Interpretability Criteria

- Clarity of explanation
- Feature importance relevance
- Ease of understanding for non-technical users

Comparison Strategy

- Each model will be evaluated using the same dataset
- Performance metrics will be compared across models
- Explainability techniques will be applied to each model
- Models will be ranked based on both performance and interpretability

The evaluation will focus not only on identifying the most accurate model but also on determining which model provides the most useful insights for real-world fraud detection.

III. Discussion

The proposed approach is expected to offer advantages over existing studies by shifting the focus from purely accuracy-based evaluation to a more holistic assessment of model usability.

While many existing approaches achieve high detection accuracy, they often fail to provide explanations that can be understood by auditors or decision-makers. This limits their practical application in real-world healthcare systems.

By incorporating interpretability into the evaluation process, the proposed framework enables a better understanding of model behavior and supports informed decision-making. This makes the system more suitable for real-world deployment, where transparency and trust are critical.

Additionally, the framework highlights the importance of balancing accuracy with interpretability, which is often overlooked in existing research. This contributes to the development of more practical and user-oriented fraud detection systems.

VII. APPLICATIONS AND USE CASES

I. Industry Applications

The proposed framework can be applied in the healthcare insurance industry to improve fraud detection and auditing processes. Insurance companies can use the framework to evaluate multiple machine learning models and identify those that provide both strong predictive performance and interpretable explanations.

The integration of SHAP and LIME explainability techniques enables auditors and investigators to understand why specific claims are classified as fraudulent, thereby improving trust and supporting faster investigation processes. This can help reduce financial losses caused by fraudulent claims while improving operational efficiency through transparent and automated decision-making.

II. Social Impact

Healthcare fraud contributes to increased insurance costs, misuse of healthcare resources, and reduced accessibility to healthcare services. By improving fraud detection accuracy and enhancing transparency through explainable AI techniques, the proposed framework can support fairer allocation of healthcare resources and reduce unnecessary financial burden on healthcare systems.

Improved transparency in fraud detection systems can also increase trust among patients, healthcare providers, insurance companies, and regulatory authorities.

III. Policy and Regulatory Relevance

Regulatory organizations increasingly require transparency, accountability, and fairness in AI-driven decision-making systems. The proposed framework supports explainable fraud detection by integrating SHAP and LIME-based interpretability techniques, which align with emerging regulatory expectations for trustworthy AI systems.

The framework can assist policymakers and healthcare organizations in developing guidelines for the adoption of interpretable and accountable AI solutions in healthcare fraud detection environments.

VIII. CONCLUSION

This study presented a comprehensive review of healthcare insurance fraud detection techniques, focusing on the evolution of methods from traditional rule-based approaches to advanced machine learning, deep learning, and hybrid models. The literature review highlighted that while existing models achieve high accuracy and improved detection capabilities, they often lack interpretability and practical usability in real-world auditing scenarios.

The analysis identified key research gaps, particularly the absence of auditor-centric explainability, the lack of systematic comparison of interpretability techniques, and the limited exploration of the trade-off between model accuracy and interpretability. These gaps indicate that current fraud detection systems are not fully aligned with the needs of real-world stakeholders, especially auditors and decision-makers.

To address these challenges, this study proposed an evaluation framework that compares multiple fraud detection models based on both performance and interpretability. The methodology emphasizes not only predictive accuracy but also the clarity and usefulness of explanations generated by the models. By incorporating explainability into the evaluation process, the proposed approach aims to enhance

transparency, trust, and usability in fraud detection systems.

The key contribution of this research lies in shifting the focus from purely performance-driven evaluation to a more holistic assessment that includes interpretability and real-world applicability. This provides a practical perspective for selecting fraud detection models that are both effective and understandable.

In conclusion, this study underscores the importance of developing fraud detection systems that balance accuracy with interpretability. Such systems are essential for improving decision-making, supporting auditors, and ensuring transparency in healthcare insurance processes. Future research can build upon this work by implementing and validating the proposed framework in real-world environments and exploring more advanced explainability techniques.

REFERENCES

[1] Nabrawi, E., & Alanazi, A. (2023). Fraud detection in healthcare insurance claims using machine learning. Risks, 11(9), 160.

[2] Dey, R., Roy, A., Akter, J., Mishra, A., & Sarkar, M. (2025). AI-driven machine learning for fraud detection and risk management in US healthcare billing and insurance. Journal of Computer Science and Technology Studies, 7(1), 188-198.

[3] Debener, J., Heinke, V., & Kriebel, J. (2023). Detecting insurance fraud using supervised and unsupervised machine learning. Journal of Risk and Insurance, 90(3), 743-768.

[4] Rukhsar, L., Bangyal, W. H., Nisar, K., & Nisar, S. (2022). Prediction of insurance fraud detection using machine learning algorithms. Mehran University Research Journal of Engineering & Technology, 41(1), 33-40.

[5] Farahmandazad, D., Danesh, K., & Abadi, H. F. N. (2025). Application of Standard Machine Learning Models for Medicare Fraud Detection with Imbalanced Data. Risks, 13(10), 198.

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

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

[8] Farbmacher, H., Löw, L., & Spindler, M. (2022). An explainable attention network for fraud detection in claims management. Journal of Econometrics, 228(2), 244-258.

[9] Saddi, V. R., Gnanapa, B., Boddu, S., & Logeshwaran, J. (2023, December). Fighting Insurance Fraud with Hybrid AI/ML Models: Discuss the Potential for Combining Approaches for Improved Insurance Fraud Detection. In 2023 4th International Conference on Communication, Computing and Industry 6.0 (C216) (pp. 01-06). IEEE.

[10] Prova, N. N. I. (2024, August). Advanced machine learning techniques for predictive analysis of health insurance. In 2024 Second International conference on intelligent cyber physical systems and internet of things (ICoICI) (pp. 1166-1170). IEEE.

[11] Ashfaq, T., Khalid, R., Yahaya, A. S., Aslam, S., Azar, A. T., Alsafari, S., & Hameed, I. A. (2022). A machine learning and blockchain based efficient fraud detection mechanism. Sensors, 22(19), 7162.

[12] Kapadiya, K., Patel, U., Gupta, R., Alshehri, M. D., Tanwar, S., Sharma, G., & Bokoro, P. N. (2022). Blockchain and AI-empowered healthcare insurance fraud detection: an analysis, architecture, and future prospects. Ieee Access, 10, 79606-79627.

[13] Amponsah, A. A., Adekoya, A. F., & Weyori, B. A. (2022). Improving the financial security of national health insurance using cloud-based blockchain technology application. International Journal of Information Management Data Insights, 2(1), 100081.

[14] Amponsah, A. A., Adekoya, A. F., & Weyori, B. A. (2022). A novel fraud detection and prevention method for healthcare claim

processing using machine learning and blockchain technology. Decision Analytics Journal, 4, 100122.

[15] Sargam, G. S., & Kalapala, R. (2025, September). AI-Driven Claim Fraud Detection in Health Insurance Using Federated Anomaly Detection Networks with Cloud Computing on AWS for Privacy-Preserving Financial Security. In 2025 Third International Conference on Cyber Physical Systems, Power Electronics and Electric Vehicles (ICPEEV) (pp. 1-6). IEEE.

[16] Lakhan, A., Mohammed, M. A., Nedoma, J., Martinek, R., Tiwari, P., Vidyarthi, A., ... & Wang, W. (2022). Federated-learning based privacy preservation and fraud-enabled blockchain IoMT system for healthcare. IEEE journal of biomedical and health informatics, 27(2), 664-672.

[17] Hasan, M. T. (2022). Graph neural network models for detecting fraudulent insurance claims in healthcare systems. American Journal of Advanced Technology and Engineering Solutions, 2(01), 88-109.

[18] Kumaraswamy, N., Markey, M. K., Ekin, T., Barner, J. C., & Rascati, K. (2022). Healthcare fraud data mining methods: a look back and look ahead. Perspectives in health information management, 19(1).

[19] Hamid, Z., Khalique, F., Mahmood, S., Daud, A., Bukhari, A., & Alshemaimri, B. (2024). Healthcare insurance fraud detection using data mining. Bmc medical informatics and decision making, 24(1), 112.

[20] Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., ... & Saif, A. (2022). Financial fraud detection based on machine learning: a systematic literature review. Applied Sciences, 12(19), 9637.

[21] Chen, Y., Zhao, C., Xu, Y., Nie, C., & Zhang, Y. (2025). Year-over-year developments in financial fraud detection via deep learning: A systematic literature review. arXiv preprint arXiv:2502.00201.

[22] Machireddy, J. R. (2023). Automation in healthcare claims processing: Enhancing efficiency and accuracy. International Journal of Science and Research Archive.

[23] Rahaman, M. M., & Islam, A. (2023). Automation And Risk Mitigation in Healthcare Claims: Policy And Compliance Implications. Review of Applied Science and Technology, 2(04), 124-157.

[24] Machireddy, J. R. (2022). Integrating predictive modeling with policy interventions to address fraud, waste, and abuse (fwa) in us healthcare systems. Advances in Computational Systems, Algorithms, and Emerging Technologies, 7(1), 35-65.

[25] Agarwal, S. (2023). An intelligent machine learning approach for fraud detection in medical claim insurance: A comprehensive study. Scholars Journal of Engineering and Technology, 11(9), 191-200.