Audit-Ready Healthcare Fraud Screening: Split-Safe Provider Aggregation and Explainable Boosted Risk Triage

Iqra Hyder, Riaz Ahmad Shaikh, Rafaqat Hussain Arain, Zahid Hussain, Basit Raza

Shah Abdul Latif University, Khairpur Mirs, Sindh, Pakistan

Abstract: Medical fraud and abnormal billing are often not clearly reflected in individual claim records, but rather in the cumulative abnormal behavior of the same service provider across multiple visits. Based on this characteristic, this paper defines the service provider, rather than a single claim, as the basic unit of risk screening and constructs a provider-level fraud screening process for auditing scenarios. Specifically, we first perform aggregation before data partitioning to minimize the risk of information leakage from the same provider across training and validation sets. Then, we train the LightGBM risk scoring model around audit-significant features such as claims volume, reimbursement, and out-of-pocket intensity, hospitalization duration statistics, duplicate claim characteristics, coding diversity, and beneficiary structure. To make the model output more suitable for actual review processes, further combine TreeSHAP interpretation, threshold scanning, and isotonic calibration, enabling the risk score to simultaneously serve priority ranking, manual review under capacity constraints, and clearer result interpretation. On the publicly available Healthcare Provider Fraud Detection dataset, based on provider-centric out-of-fold evaluation, the proposed method achieves good ranking performance, with an AUC of 0.939, an AUPRC of 0.699, and an F1 score of 0.666 at the selected threshold. The results also show that maximum hospitalization duration, reimbursement intensity, total claims volume, total out-of-pocket expenses, and beneficiary age structure are key risk signals. This provides a dense, interpretable, and auditable enactment method for beneficiary risk showing in health claims settings.

Keywords: Healthcare Fraud Detection, Provider-Level Screening, Explainable Machine Learning, Lightgbm, Treeshap, Risk Triage.

Email: iqrahyder.cs42@gmail.com

1. Introduction

The healthcare organization makes a large number of due records daily; nonetheless, the Special Investigations Unit (SIU) can only physically review a very limited number of cases. For health scams, the problem is frequently not with a single inaccessible prerogative but rather with the irregular structures collected from manifold visits to the same service provider, such as irregular hospital vacation period, reliably high repayment levels, or repeated submissions. In actual auditing and screening work, using the service provider rather than a single claim as the unit of analysis and decision making is often more in line with the real business process [1] – [7]. Latest systematic reviews also point out that while research on medical fraud detection continues to increase, there are still important inadequacies in this field concerning label acquisition,

standardized experimental protocols, interpretability of results, and connection with real operational processes [6],[7].

Current methods have slowly evolved from arithmetical investigation and regulation systems to supervised learning, cost-sensitive learning, and graph structure modeling [1]-[15]. However, for tasks aimed at generating SIU review queues, model performance is not the only objective: if training and validation are randomly divided at the claims level, multiple records of the same provider may appear in different data subsets at the same time, resulting in information leakage and overestimation of effectiveness. At the same time, the screening system also needs to provide explainable reasons for high risk and support thresholding output based on audit capacity. Motivated by this practical need, this research proposes a compact provider-level risk screening framework: split-safe provider-level aggregation is completed before the division, and the LightGBM risk model is trained with aggregated features that have audit significance. Together with TreeSHAP, threshold sweeping, and post-hoc calibration, the model score is transformed into provider ranking results that are more suitable for manual review and operational decision-making [8], [9],[16],[17],[18]. Unlike prior work, it emphasizes provider-level split safety, audit-meaningful features, and capacity-aware operational triage.

As shown in Fig. 1, the proposed framework first constructs provider-level features from claims data, then performs split-safe validation and LightGBM-based risk scoring, and lastly supports operational triage through explanation, threshold selection, and calibration.

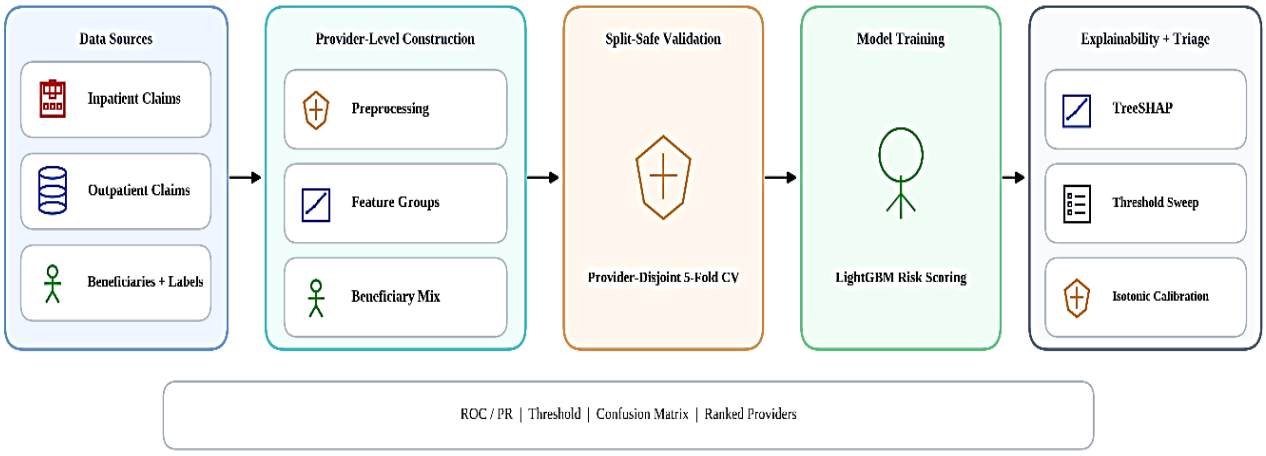

Fig. 1 System Overview of the Proposed Split-Safe, Provider-Level Healthcare Fraud Screening Pipeline

2. Related Work

Early research on medical fraud detection was mainly based on statistical anomaly detection, rule discovery, and data mining methods. Bolton and Hand summarized fraud identification as a special learning problem that simultaneously faces class imbalance, behavioral evolution, and high false alarm costs [1]. Li et al. further systematically summarized statistical methods in medical fraud detection and pointed out that medical service data has typical characteristics such as multi-subject interaction, complex business rules, and insufficient labels [2]. Subsequently, Joudaki et al. and Bauder et al. further discussed provider-level fraud, abuse, and upcoding from an operational auditing perspective that anomaly identification at the provider level not only has research value but also directly corresponds to the actual operational unit in medical insurance payment auditing [3]–[5]. In recent years, supervised learning has gradually become mainstream. Random forests, gradient boosting, cost-sensitive learning, and other tabular models remain competitive on structured medical claims data, especially suitable for structured, heterogeneous, but relatively interpretable feature spaces [6], [8], [9], [12], [15].

Meanwhile, relational modeling and graph neural network methods have developed rapidly in recent years, with researchers using heterogeneous connections between patients, providers, institutions, and services to improve the ability to characterize complex fraud patterns [10]–[14]. These methods have advantages in expressing relational structures, but are usually accompanied by more complex data preparation, graph construction, and deployment costs. In contrast, tree models based on provider-level aggregated features still have a better performance-interpretability-implementation cost balance in many operational scenarios. LightGBM provides a mature tool for efficient table learning, while TreeSHAP supports global driver analysis and individual-level cause explanation [16], [17]. Operational screening studies in neighboring domains have shown that discriminative ability alone is not enough to support reliable decision-making, and calibrated probabilities and uncertainty-aware outputs are also important for high-throughput screening tasks [18]. Therefore, it does not pursue new graph architectures or complex deep models, but emphasizes integrating provider-level split-safe aggregation, auditable table features, LightGBM risk scoring, TreeSHAP interpretation, and capacity-aware thresholding output into a practical framework for SIU workflows.

2.1 Data and Provider-Level Feature Construction

The Healthcare Provider Fraud Detection Analysis dataset is used in this research. The original dataset mainly comprises inpatient claim forms, outpatient claim forms, recipient info slabs, and provider label tables. The training portion includes provider risk labels, while the testing portion includes unlabeled providers. Considering that the decision-making object for actual auditing and review is the service provider rather than a single claim, this paper uniformly uses the provider as the prediction unit and completes data connection, aggregation, and subsequent modeling around this unit.

In the data preprocessing stage, the identifier fields such as Provider, BeneID, and ClaimID are initially read as strings to evade the damage of leading zeros or misunderstanding of identifier information. Date fields are normalized, and monetary fields such as reimbursement amount and deductible amount are converted to numeric types. The beneficiary information table is redid and prepared, retaining key attributes such as age, region, and chronic disease markers, and left-joined with claim records using BeneID. For hospitalization claims, the length of stay (LOS) is further calculated based on admission and discharge dates; for duplicate claims, the number of occurrences is totaled within the provider based on the standardized beneficiary start and end date combinations. To minimalize information leakage caused by the same provider appearing in both the training and validation stages, all provider-level features are aggregated before data partitioning.

In terms of feature creation, it retains only provider-level facts and figures that are directly relevant to audit judgments and easy to understand. These includes utilization features reflecting the scale of service use, such as inpatient, outpatient, and total claim volumes, as well as the number of sole beneficiaries; financial strength features characterizing the level of fund consumption, such as the sum, mean, and maximum of compensation and deductible expenditures; LOS features describing inpatient behavior patterns, such as the mean, median, and maximum length of hospital stay; coding pattern features reflecting coding complexity, including code density and code diversity; duplication intensity features used to identify duplicate claims; and beneficiary mix features summarizing the service recipient structure, including age facts and figures, the amount of chronic diseases, the relation of inpatient to outpatient visits, and local entropy. This feature design maintains the compactness and interpretability of the input space and is largely consistent with the more important risk signals in the subsequent results, especially

the maximum length of stay (LOS), reimbursement strength, claims volume, total out-of-pocket expenses, and beneficiary age structure.

3. Methodology

3.1 Split-Safe Validation

To avoid the same service provider appearing in both the training and validation sets, this paper employs provider-level disjoint partitioning, i.e., cross-validation is performed by provider rather than by individual claims. All claims cleansing, beneficiary linking, and provider-level feature aggregation are completed before partitioning, but training and validation are always bounded by different provider sets, thereby reducing evaluation bias caused by information leakage. Employments a five-fold stratified cross-validation approach, checking for provider non-overlap, relatively stable label ratios, and consistent feature column order within each fold to ensure the validation results more accurately reflect the screening effectiveness on unseen providers.

3.2 Model Training

At the modeling stage, we use LightGBM to assign provider-level risk scores. Considering the imbalanced distribution of positive and negative samples, class weights are used during training to mitigate the problem of minority classes being ignored, and conventional regularization and early stopping strategies are combined to control overfitting. The model output is a risk score for each provider; within the cross-validation framework, the out-of-fold (OOF) predictions for each trained provider are further retained for subsequent unified calculation of ROC, PR, threshold scan, and confusion matrix. Thus, the AUC, AUPRC, F1, and working thresholds given later are all based on OOF predictions, rather than optimistic same-set evaluations.

3.3 Explainability and Operationalization

To improve the interpretability and usability of the model output, adds three steps to the final workflow: TreeSHAP, threshold scanning, and isotonic calibration. First, TreeSHAP outputs global feature drivers and individual-level explanation outputs, facilitating the explanation of why a particular provider is marked as high-risk. Numerous applicant verges are skimmed centered on the OOF score, comparing accuracy, recall, and F1 at dissimilar thresholds to choose an additional suitable action fact built on the monthly appraisal volume of SIU. Last of all, an isotonic method is rummage-sale for post-hoc calibration of the original scores, making the risk value closer to an interpretable probability expression within the operation interval? The model not only ranks providers for risk, but also mutual with verge sceneries, making a showing list

more suitable to review capacity, though making the understanding of risk scores laid-back for applied auditing usage.

4. Results and Discussion

Based on provider-level out-of-fold (OOF) prediction results, LightGBM demonstrates strong discriminative ability in the risk ranking task. Looking at the overall curves, the ROC curve remains at a high level, and the PR curve is significantly higher than the positive class baseline, indicating that the model can not only distinguish between high-risk and low-risk providers well, but also has good screening value under conditions of imbalanced class distribution. Further, from a quantitative perspective, the model achieved an AUC of 0.939 and an AUPRC of 0.699, indicating that the proposed method has relatively robust performance in the provider-level risk ranking task on this public dataset.

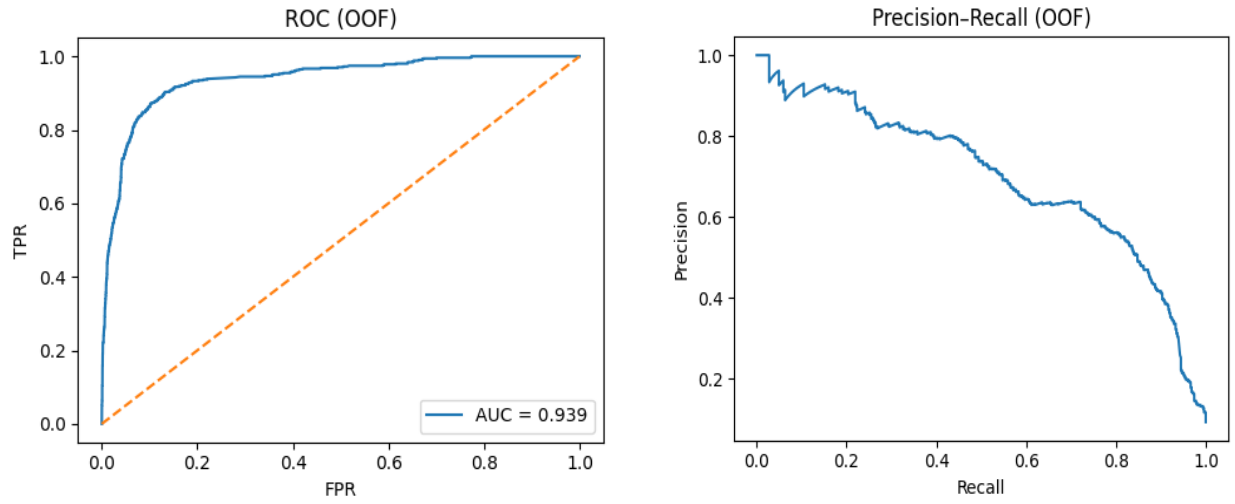

Fig. 2. OOF Discrimination Performance of the proposed model: (a) ROC curve and (b) precision–recall curve

Combined with threshold scanning results, it selects $\tau_{CV}=0.425$ as the operating threshold for cross-validation. At this threshold, the model's F1 score is 0.666, achieving a relatively balanced trade-off between recall and precision. The corresponding confusion matrix shows that the model identifies a large proportion of high-risk providers while keeping the review queue at a manageable scale, making it more suitable as a pre-screening tool for SIU review queues. Threshold scanning also indicates that lower thresholds are more conducive to improving recall, while higher thresholds are more conducive to improving precision, providing a direct basis for strategy selection under different audit capacities.

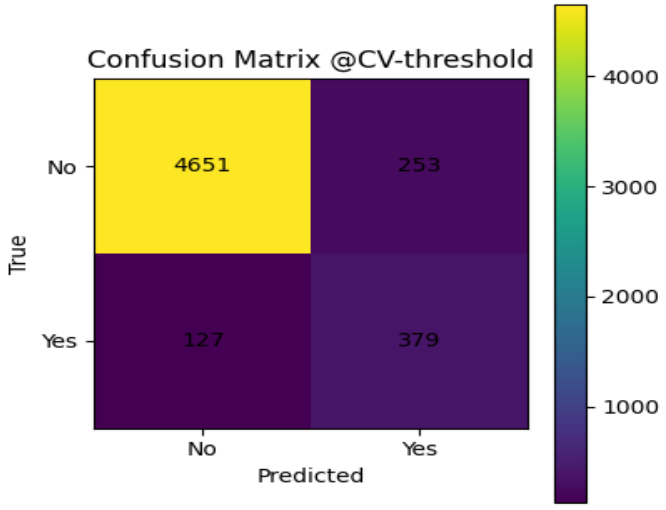

Fig. 3 Confusion Matrix at the Selected Operating Threshold τCV

From the interpretation results, the most important risk drivers of the model mainly include the maximum length of hospital stay, inpatient and outpatient reimbursement intensity, total claims, inpatient out-of-pocket expenses, and beneficiary age structure. These signals are largely consistent with common audit concerns in medical fraud or abnormal billing: on the one hand, abnormally high length of stay (LOS) and reimbursement amounts may correspond to a persistently high-cost service model; on the other hand, large claims volume, high out-of-pocket expenses, and specific beneficiary structures may also jointly drive up provider risk scores. These results indicate that the provider-level aggregated features used in this paper not only have statistical discriminative power but also retain good business interpretability.

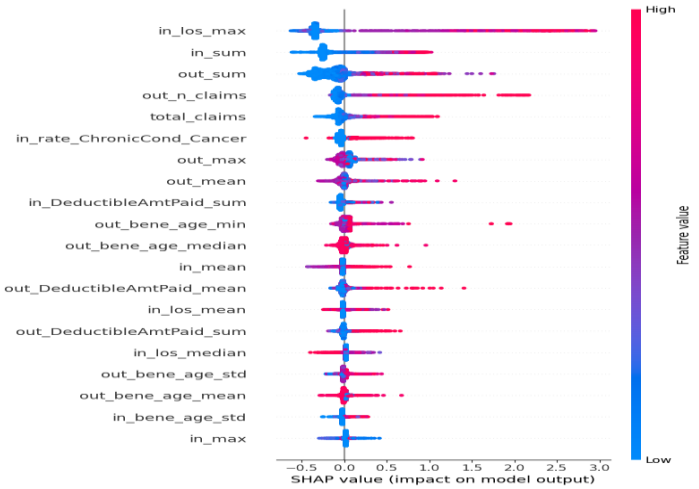

Fig. 4. SHAP Summary Showing the Impact of Key Provider-Level Features on Model Output

At the probabilistic level, the isotonic-calibrated reliability curve shows reasonable alignment with the diagonal over the main operating range inside the focal functioning range, indicating that the model output score has better readability in review ranking and risk stratification scenarios. In summary, this is not intended to replace manual investigation but is more suitable as a provider-level, interpretable, capacity-aware risk screening process: first, the model completes the high-risk ranking, and then, combined with thresholds and explanation outputs, generates an auditable review list, thereby providing more targeted priority support for the limited manpower of SIUs.

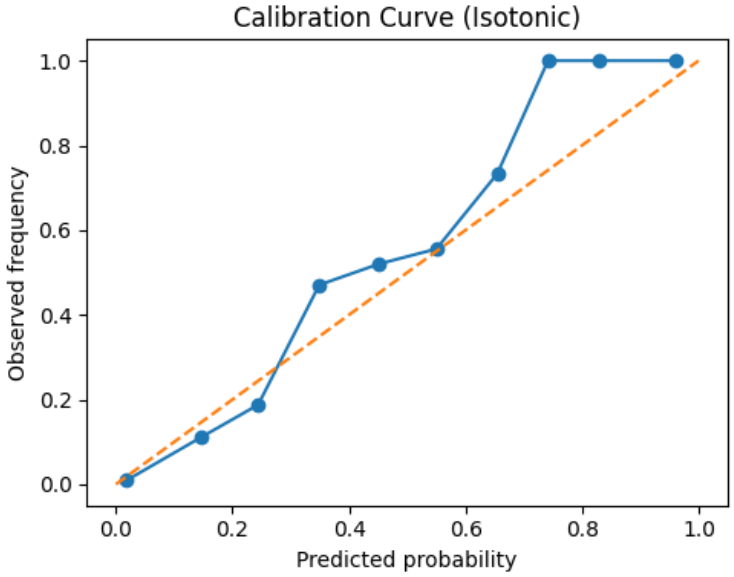

Fig. 5. Reliability plot after isotonic calibration

5. Conclusion

This paper proposes a provider-level fraud risk screening process for medical claims auditing scenarios. It uses service providers as a unified prediction unit, performs provider-level aggregation before data partitioning, and reduces the risk of information leakage through provider mutual exclusion verification. Based on this, LightGBM is used to score the risk of aggregated features with audit implications, and combined with TreeSHAP, threshold scanning, and isotonic calibration, the model output is transformed into provider ranking results more suitable for manual review and capacity constraint management.

In provider-level OOF evaluations on public datasets, we achieved good ranking performance, with an AUC of 0.939, an AUPRC of 0.699, and an F1 score of 0.666 near the selected

operational threshold of 0.425. The results demonstrate that all-out length of hospital stay, reimbursement intensity, total claims, deductible expenses, and beneficiary structure are key factors. Indicating that even using lone compact and audit provider-level aggregated structures, the tree model yet provides practically valuable support for high-risk provider checking.

The significance of this learning is not to substitute manual examination, but to bring SIUs a more easily implemented pre-screening framework. Concluded threshold output result understanding and calibrated risk scores, auditors can more efficiently timetable appraisal arrangements with incomplete manpower, forming a clearer and more defensible work queue. Consequently, a more suitable understanding is that it proposes a provider-level risk triage arrangement for practical review processes rather than a concluding response to the interrogation of health fraud recognition.

Some obvious limitations are still there. The experiments are currently conducted on a single public dataset, and the model's generalization ability in other data settings requires more validation. The prevailing labels themselves may not be comprehensive, and some newer, more delicate, irregular behaviors may not be completely characterized by the current aggregated features. The essential process does not yet incorporate graph structure relationship information, weak supervision mechanisms, or more fine-grained fairness analysis, thus its coverage of collaborative behavior and complex network patterns remains limited.

Several directions are available for future work. The model's constancy essentials to be verified on better data sources and dissimilar duration portions, and the impact of circulation drift needs to be constantly evaluated. Graph features, weak supervision, or richer relationship information can be supplementary introduced to recover the ability to classify clique-related fraud and multifaceted communication configurations. Research on group calibration, fairness testing, and uncertainty control can also aid in improving the stability and interpretability of this category of provider-level risk screening methods in real-world audit scenarios. The results demonstrate that a provider-centric modeling tactic that emphasizes split safe verification and incorporates interpretability mechanisms can provide a compact and promising basic solution for SIU risk triage in medical claims scenarios.

References

[1]. R. J. Bolton and D. J. Hand, "Statistical Fraud Detection: A Review," *Statistical Science*, vol. 17, no. 3, pp. 235–255, 2002, doi: 10.1214/ss/1042727940.

[2]. J. Li, K.-Y. Huang, J. Jin, and J. Shi, "A Survey on Statistical Methods for Health Care Fraud Detection," *Health Care Management Science*, vol. 11, no. 3, pp. 275–287, 2008, doi: 10.1007/s10729-007-9045-4.

[3]. H. Joudaki, A. Rashidian, B. Minaei-Bidgoli, M. Mahmoodi, B. Geraili, M. Nasiri, and M. Arab, "Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature," *Global Journal of Health Science*, vol. 7, no. 1, pp. 194–202, 2014, doi: 10.5539/gjhs.v7n1p194.

[4]. R. Bauder, T. M. Khoshgoftaar, and N. Seliya, "A Survey on the State of Healthcare Upcoding Fraud Analysis and Detection," *Health Services and Outcomes Research Methodology*, vol. 17, no. 1, pp. 31–55, 2017, doi: 10.1007/s10742-016-0154-8.

[5]. H. Joudaki, A. Rashidian, B. Minaei-Bidgoli, M. Mahmoodi, B. Geraili, M. Nasiri, and M. Arab, "Improving Fraud and Abuse Detection in General Physician Claims: A Data Mining Study," *International Journal of Health Policy and Management*, vol. 5, no. 3, pp. 165–172, 2016, doi: 10.15171/ijhpm.2015.196.

[6]. A. du Preez, S. Bhattacharya, P. Beling, and E. Bowen, "Fraud Detection in Healthcare Claims Using Machine Learning: A Systematic Review," *Artificial Intelligence in Medicine*, vol. 160, art. no. 103061, 2025, doi: 10.1016/j.artmed.2024.103061.

[7]. A. V. Najar, L. Alizamani, M. Zarqi, and E. Hooshmand, "A Global Scoping Review on the Patterns of Medical Fraud and Abuse: Integrating Data-Driven Detection, Prevention, and Legal Responses," *Archives of Public Health*, vol. 83, no. 1, art. no. 43, 2025, doi: 10.1186/s13690-025-01512-8.

[8]. Z. Hamid, F. Khalique, S. Mahmood, A. Daud, A. Bukhari, B. Alshemaimri, *et al*., "Healthcare Insurance Fraud Detection Using Data Mining," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, art. no. 112, 2024, doi: 10.1186/s12911-024-02512-4.

[9]. E. Nabrawi and A. Alanazi, "Fraud Detection in Healthcare Insurance Claims Using Machine Learning," *Risks*, vol. 11, no. 9, art. no. 160, 2023, doi: 10.3390/risks11090160.

[10]. J. Lu, K. Lin, R. Chen, M. Lin, X. Chen, and P. Lu, "Health Insurance Fraud Detection by Using an Attributed Heterogeneous Information Network with a Hierarchical Attention Mechanism," *BMC Medical Informatics and Decision Making*, vol. 23, art. no. 62, 2023, doi: 10.1186/s12911-023-02152-0.

[11]. B. Hong, P. Lu, H. Xu, J. Lu, K. Lin, and F. Yang, "Health Insurance Fraud Detection Based on Multi-Channel Heterogeneous Graph Structure Learning," *Heliyon*, vol. 10, no. 9, art. no. e30045, 2024, doi: 10.1016/j.heliyon.2024.e30045.

[12]. N. Kumaraswamy, T. Ekin, C. Park, M. K. Markey, J. C. Barner, and K. Rascati, "Using a Bayesian Belief Network to Detect Healthcare Fraud," *Expert Systems with Applications*, vol. 238, art. No. 122241, 2024, doi: 10.1016/j.eswa.2023.122241.

[13]. S. Mardani and H. Moradi, "Using Graph Attention Networks in Healthcare Provider Fraud Detection," *IEEE Access*, vol. 12, pp. 132786–132800, 2024, doi: 10.1109/ACCESS.2024.3425892.

[14]. R. Muhammad, D. Tbaishat, A. Nazir, S. Yacoub, M. A. A. AbdulRazek, M. A. A. El-Enen, and A. T. Sahlol, "Fraud Detection and Explanation in Medical Claims Using GNN Architectures," *Scientific Reports*, vol. 15, art. no. 41734, 2025, doi: 10.1038/s41598-025-22910-6.

[15]. H. Shi, M. A. Tayebi, J. Pei, and J. Cao, "Cost-Sensitive Learning for Medical Insurance Fraud Detection With Temporal Information," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 10, pp. 10451–10463, 2023, doi: 10.1109/TKDE.2023.3240431.

[16]. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 2017, pp. 3146–3154, doi: 10.5555/3294996.3295074.

[17]. S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, *et al*., "From Local Explanations to Global Understanding with Explainable AI for Trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.

[18]. B. Raza, A. Maitlo, Z. H. Shar, and I. Hyder, "Operational Android Malware Filtering: Calibrated Probabilities and Distribution-Free Guarantees," *Kashf Journal of Multidisciplinary Research*, vol. 2, no. 12, pp. 58–73, 2025, doi: 10.71146/kjmr778.