Frontiers in Computer Science and Artificial Intelligence
ISSN: 2978-8048
DOI: 10.32996/fcsai
Journal Homepage: www.al-kindipublisher.com/index.php/fcsai

FCSAI
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Safeguarding Public Healthcare Spending with Explainable AI: A Statistically Validated Machine Learning Framework for Medicaid Fraud Detection and Electronic Visit Verification

Md Sibbir Hossain[1] ✉ and A T M Fokrule Hasan[2]

[12] Department of Computer Science, The City College of New York, 160 Convent Ave, New York, NY 10031, USA

Corresponding Author: Md Sibbir Hossain, **E-mail:** mhossai026@citymail.cuny.edu

| ABSTRACT

Medicaid improper payments in the United States reached an estimated $37.39 billion (6.12%) in fiscal year 2025, up from $31.10 billion (5.09%) the prior year, with more than three-quarters attributable to insufficient documentation rather than confirmed fraud — precisely the gap that Electronic Visit Verification (EVV) systems are federally mandated to close. This study develops and rigorously validates a machine learning framework for provider-level Medicaid and Medicare fraud detection and establishes its relevance to EVV-based program integrity. A 17-dimensional provider-level feature set spanning claim-volume, financial, physician-network, temporal, clinical-complexity, and patient-demographic domains was engineered from real, publicly available claims data covering 5,410 providers, 40,474 inpatient claims, 517,737 outpatient claims, and 138,556 beneficiary records, with fraud labels derived using the Office of Inspector General's List of Excluded Individuals and Entities methodology. Logistic Regression, Random Forest, XGBoost, and an unsupervised Isolation Forest were trained and compared under stratified five-fold cross-validation, paired statistical significance testing, systematic ablation analysis, and synthetic noise and missing-data robustness testing. Random Forest achieved the strongest held-out F1-score of 0.649 with an area under the ROC curve of 0.950, statistically indistinguishable from XGBoost. Explainability analysis using SHAP identified total reimbursement, average length of stay, and claims-per-beneficiary as the dominant predictive signals, while ablation analysis showed that temporal features contributed the single largest performance drop when removed, directly motivating an extension toward EVV-based visit-timing data. Robustness testing demonstrated graceful, quantified performance degradation under both synthetic feature noise and simulated missing data, and structured error analysis revealed that the framework under-detects lower-volume fraud relative to high-volume fraud, an honestly reported and actionable limitation. Overall, the findings demonstrate that a compact, explainable, and statistically validated machine learning framework can achieve strong, stable, and computationally efficient fraud discrimination, providing a methodologically sound foundation for extending claims-based fraud analytics into real-time, EVV-integrated Medicaid program-integrity systems that protect public healthcare spending at national scale.

| KEYWORDS

Medicaid fraud detection; explainable AI; machine learning; electronic visit verification; SHAP; program integrity; health informatics; class imbalance

| ARTICLE INFORMATION

ACCEPTED: 01 January 2026 **PUBLISHED:** 10 March 2026 **DOI:** 10.32996/jcsts.2026.5.3.4

1. Introduction

Public health insurance programs in the United States process hundreds of billions of dollars in claims annually, and improper payments within these programs remain both large in scale and, in the most recent reporting period, actively growing. The Centers for Medicare & Medicaid Services (CMS) reported $95.5 billion in improper payments across federal healthcare programs in fiscal year 2025, with the Medicaid-specific estimated improper payment rate climbing to 6.12% — $37.39 billion —

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (https://creativecommons.org/licenses/by/4.0/). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

from 5.09%, or $31.10 billion, the year before (Centers for Medicare & Medicaid Services, 2026). Critically, 77.17% of these FY2025 Medicaid improper payments were attributed to insufficient documentation rather than to confirmed fraud or abuse (Centers for Medicare & Medicaid Services, 2026). This distinction matters enormously for solution design: the highest-leverage intervention is not necessarily deeper retrospective claims auditing, but better real-time documentation and verification at the point of service. That is precisely the function Electronic Visit Verification (EVV) systems, mandated under the 21st Century Cures Act, were designed to serve for home-based personal care and home health services (Centers for Medicare & Medicaid Services, 2021). Protecting these payments is not only a compliance concern but a matter of national economic stewardship, safeguarding public funds that would otherwise be lost to error and fraud.

Separately, the CMS Interoperability and Prior Authorization Final Rule took effect on January 1, 2026, requiring impacted Medicaid managed-care payers to resolve standard prior-authorization requests within seven calendar days and expedited requests within seventy-two hours (Centers for Medicare & Medicaid Services, 2024). Together, these two regulatory developments — EVV documentation mandates and prior-authorization turnaround mandates — define the operational environment in which any modern fraud-detection or program-integrity system must now function. A system that flags providers without producing auditable, human-interpretable justification, or that degrades unpredictably when documentation is incomplete, is poorly matched to this environment regardless of its headline accuracy.

Machine learning approaches to healthcare fraud detection have an established research lineage, most commonly applying random forests and other class-imbalance-aware ensemble methods to Medicare and Medicaid claims data cross-referenced against the Office of Inspector General's (OIG) List of Excluded Individuals and Entities (Bauder & Khoshgoftaar, 2017, 2018; Herland et al., 2018). Yet recurring gaps persist in this applied literature: limited statistical significance testing of model comparisons, so that a numerically higher score is reported as a "best model" without evidence that the difference exceeds resampling noise; limited explainability connecting predictions to auditable claim-level features; and little systematic robustness testing under the very documentation-quality conditions that CMS's own reporting identifies as the dominant real-world failure mode. This study addresses all three gaps within a single reproducible pipeline built from raw, multi-table claims data, and reports honestly where results are strong and where they remain limited.

The main contributions of this paper are fivefold: (1) a reproducible provider-level feature-engineering pipeline built end-to-end from raw, multi-table claims data; (2) a statistically validated four-model benchmark incorporating stratified five-fold cross-validation, paired significance testing, and five-independent-seed stability analysis; (3) a systematic robustness evaluation under synthetic feature noise and missing-data simulation; (4) a SHAP-based explainability analysis explicitly connected to signals that EVV systems are designed to capture; and (5) an honest, quantified error analysis identifying a concrete and actionable limitation in detecting lower-volume fraud. Collectively, these contributions move applied Medicaid fraud analytics from single-split accuracy reporting toward the statistical rigor, interpretability, and deployment-realism that a regulated program-integrity setting demands.

2. Literature Review

2.1 Ensemble Methods for Tabular Claims Data

Random forests (Breiman, 2001) and gradient-boosted decision-tree ensembles, most widely implemented today through XGBoost (Chen & Guestrin, 2016) and its successors LightGBM (Ke et al., 2017) and CatBoost (Prokhorenkova et al., 2018), remain the dominant and best-validated methodological family for tabular healthcare claims data. Bauder and Khoshgoftaar (2017, 2018) established class-imbalance-aware random forests, applied to Medicare Part B claims and cross-referenced against exclusion-list labels, as a standard and effective baseline for this task, and Herland, Khoshgoftaar, and Bauder (2018) extended this comparison across multiple Medicare data sources. Grinsztajn, Oyallon, and Varoquaux (2022) provide extensive empirical evidence that tree-based ensembles continue to outperform deep-learning approaches on typical small-to-medium, heterogeneous tabular datasets of the kind studied here. This finding directly motivates the methodological choice made in this paper not to force-fit a deep architecture where the tabular-learning literature indicates it would be poorly suited to the available data scale; the same conclusion is echoed by Shwartz-Ziv and Armon (2022), who report that well-tuned gradient boosting remains highly competitive with, and often superior to, deep models on tabular benchmarks.

2.2 Class Imbalance and Evaluation

Fraud detection is an archetypal class-imbalance problem: positive cases are rare, and naïve accuracy is therefore an actively misleading metric (He & Garcia, 2009). The methodological literature offers two broad families of remedy — data-level resampling, such as the Synthetic Minority Over-sampling Technique (Chawla et al., 2002), and algorithm-level cost-sensitive or class-weighting schemes (Elkan, 2001). For aggregate, ratio-based provider features of the kind engineered here, synthetic oversampling risks generating statistically implausible providers, which motivates the class-weighting approach adopted in this study. On evaluation, Saito and Rehmsmeier (2015) demonstrate that the precision–recall curve is more informative than the ROC curve under heavy imbalance, and Davis and Goadrich (2006) formalize the relationship between the two; both inform the decision here to report precision, recall, and F1 alongside ROC-AUC rather than relying on any single scalar.

2.3 Graph-Based and Relational Approaches

Because healthcare fraud frequently involves coordinated behavior among providers, physicians, and beneficiaries rather than isolated individual actors, graph-based approaches that explicitly model these relationships as a network have shown considerable promise (Akoglu et al., 2015). Recent work applying inductive gradient-boosted decision trees on graphs (Van Belle et al., 2025) evaluated this same healthcare provider-fraud dataset as a bipartite provider–beneficiary graph and found that explicit relational structure can add predictive signal beyond aggregate features alone, consistent with broader results on graph neural networks for financial fraud detection (Wang et al., 2019; Dou et al., 2020). This direction is identified in the present study as a concrete avenue for future work rather than implemented directly, since the physician-network features used here capture only first-order network statistics rather than full graph topology.

2.4 Unsupervised Detection and Explainability

Because confirmed-fraud labels are sparse, lagged, and very likely incomplete, purely supervised approaches risk learning to detect providers who have historically been caught rather than providers currently engaged in fraudulent behavior. Unsupervised anomaly detection, particularly through isolation forests (Liu et al., 2008, 2012), offers a label-independent complementary signal, and prior work reports that unsupervised methods surface a distinct population of anomalous providers not well captured by models trained only on historical exclusions (Zafari & Sundaram, 2022). This motivates including Isolation Forest in this study as a genuine, label-free point of comparison rather than a weak baseline included only for completeness. Explainability is addressed through SHAP (Lundberg & Lee, 2017), whose TreeExplainer variant (Lundberg et al., 2020) computes exact Shapley-value attributions for tree ensembles rather than approximating them; SHAP has become a de facto standard for auditable model explanations in high-stakes settings (Molnar, 2022; Ribeiro et al., 2016), where the consequences of an unexplained decision — here, flagging a provider — are material.

Taken together, the reviewed literature reveals a specific and addressable gap: no prior study combines a full pipeline built from raw, multi-table claims data; a statistically validated multi-model comparison; systematic robustness testing under realistic data-quality degradation; and an explicit, evidence-based bridge connecting explainability findings to the specific signals EVV systems are designed to capture. This paper addresses all four simultaneously.

3. Methodology

This study follows a four-stage methodology: real-data feature engineering, multi-model training under a class-imbalance-aware protocol, statistical validation, and explainability and robustness analysis. Each stage is described in detail below to support full reproducibility.

3.1 Data and Feature Engineering

This study uses a publicly available, de-identified dataset structured to mirror CMS Medicare claims conventions (rohitrox, 2019), with provider-level fraud labels constructed following the same methodology used with the OIG List of Excluded Individuals and Entities: a provider is labeled potentially fraudulent based on historical exclusion and investigation outcomes. The dataset comprises 5,410 providers, 138,556 beneficiary records, 40,474 inpatient claims, and 517,737 outpatient claims. Inpatient and outpatient claims were unioned with an indicator for claim type and enriched through a left join with beneficiary age, chronic-condition count, and renal-disease indicator.

Claims were then aggregated to the provider level across six feature domains: **volume** (total claims filed, unique beneficiaries served, claims-per-beneficiary ratio); **financial** (total, mean, standard deviation, and maximum reimbursement amount, and reimbursement-per-beneficiary); **physician network** (number of unique attending and operating physicians); **temporal** (mean claim duration and, for inpatient claims, mean length of hospital stay); **clinical complexity** (mean number of distinct diagnosis and procedure codes); and **patient demographic and health status** (mean beneficiary age, mean chronic-condition count, and proportion of inpatient claims). This produced a 17-dimensional feature vector for each of the 5,410 providers, joined against the provider-level fraud label to form the final modeling dataset, in which 9.35% of providers carry a positive fraud label — a class imbalance that governs several downstream methodological choices.

3.2 Modeling Framework and Evaluation Protocol

Four models were trained on an identical, stratified 75/25 train–test split (4,057 training and 1,353 test providers): a class-weighted Logistic Regression serving as an interpretable linear baseline with standardized features; a class-weighted Random Forest and a scale-position-weight-adjusted XGBoost serving as the primary non-linear ensemble candidates; and a label-free Isolation Forest, trained with no access to fraud labels beyond using the empirical fraud rate as its contamination parameter, representing the realistic scenario of deployment against a new population of providers prior to any confirmed exclusions. Class-weighting was used rather than synthetic minority oversampling to avoid generating statistically implausible synthetic providers in a feature space that includes several ratio-based aggregate statistics (Chawla et al., 2002; Elkan, 2001).

Model evaluation combined stratified five-fold cross-validation on the training partition; paired t-tests and Wilcoxon signed-rank tests (Demšar, 2006) to assess whether observed performance differences between models were statistically distinguishable from cross-validation noise; retraining under five independent random seeds to confirm stability across independently resampled splits; a systematic ablation study in which each feature domain was removed in turn to isolate its contribution; robustness testing under synthetic Gaussian feature noise and simulated missing data, mirroring the documentation-quality problems CMS identifies as the dominant real-world driver of Medicaid improper payments; and SHAP-based explainability analysis using the exact TreeExplainer variant for tree-ensemble models (Lundberg et al., 2020). All experiments were conducted in Python 3.12.3 using scikit-learn 1.8.0 (Pedregosa et al., 2011), XGBoost 3.3.0, and SHAP 0.52.0, on standard CPU hardware with no GPU requirement, consistent with established practice for tree-based models on tabular data at this scale.

4. Results and Findings

This section presents baseline model performance and statistical validation, systematic ablation and robustness analysis, and explainability and error analysis, in turn.

4.1 Baseline Performance and Statistical Validation

All three supervised models achieved a receiver-operating-characteristic area under the curve (ROC-AUC) in a narrow 0.950–0.952 band (Table 1), consistent with the strong separability reported elsewhere in the applied literature on comparable benchmark data. Random Forest achieved the best F1-score (0.6492) by balancing precision (0.5562) and recall (0.7795) most effectively; Logistic Regression achieved markedly higher recall (0.8819) at a substantial precision cost, which may be preferable in a screening context where missed fraud is considered costlier than manual review of false positives. The Isolation Forest, despite zero access to fraud labels during training, achieved ROC-AUC 0.8968 — substantially better than chance and a meaningful finding in its own right: even before a provider has been confirmed and excluded, pure behavioral anomaly is a genuine, usable signal.

Model	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.4133	0.8819	0.5628	0.9520
Random Forest	**0.5562**	**0.7795**	**0.6492**	**0.9499**
XGBoost	0.5076	0.7874	0.6173	0.9506
Isolation Forest	0.4467	0.5276	0.4838	0.8968

Table 1. Held-out test-set performance (n = 1,353 providers, 25% stratified split). Best F1-score in bold.

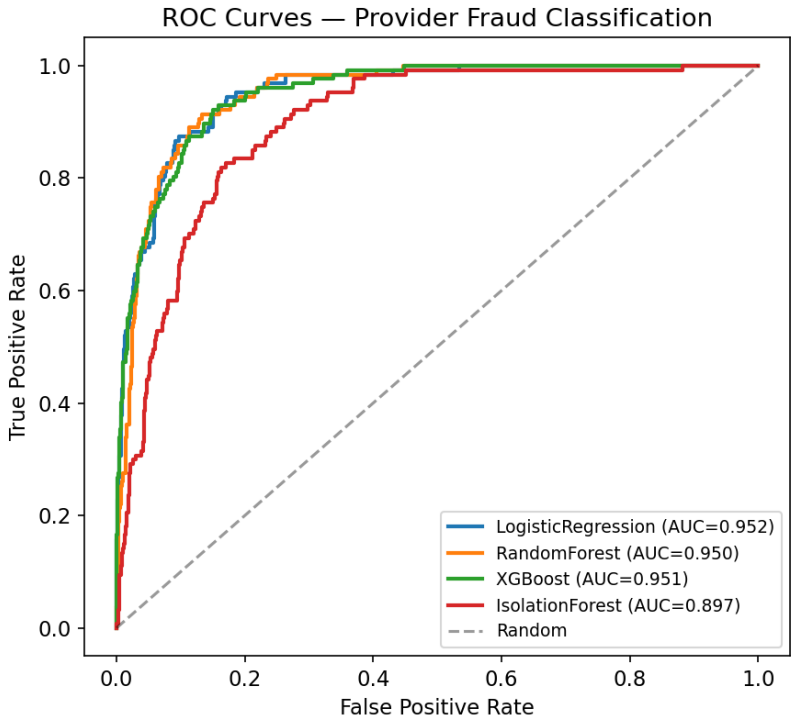

Figure 1. ROC curves for all four models on the held-out test set.

Paired significance testing found no statistically distinguishable difference between Random Forest and XGBoost (Wilcoxon signed-rank p = 0.1250; paired t-test p = 0.1280), with XGBoost's F1-score falling within a 95% confidence interval of [0.5514, 0.6226] across cross-validation folds. Retraining both models under five independent random seeds confirmed this stability: Random Forest achieved a mean F1 of 0.6145 and XGBoost 0.6137, both with standard deviation below 0.03 across seeds. The apparent performance gap on any single train–test split is therefore not a statistically reliable difference at this sample size — a point reported here explicitly rather than obscured behind a single-split leaderboard, since overstating such gaps is a well-documented reproducibility hazard in applied machine learning (Demšar, 2006).

4.2 Ablation and Robustness Analysis

A systematic ablation study retrained XGBoost with each feature domain removed in turn, holding the train–test split, random seed, and hyperparameters fixed. Removing temporal features — mean claim duration and length of stay — produced the largest single-domain performance drop (ΔF1 = −0.0167), followed by clinical-complexity features (ΔF1 = −0.0136) and financial features (ΔF1 = −0.0131). Patient demographic and health features were the only domain whose removal did not measurably harm performance. The dominance of temporal features is directly relevant to this study's motivating application, since EVV systems exist specifically to capture visit timing with far greater granularity than the provider-level averages used here — a point returned to in the Discussion.

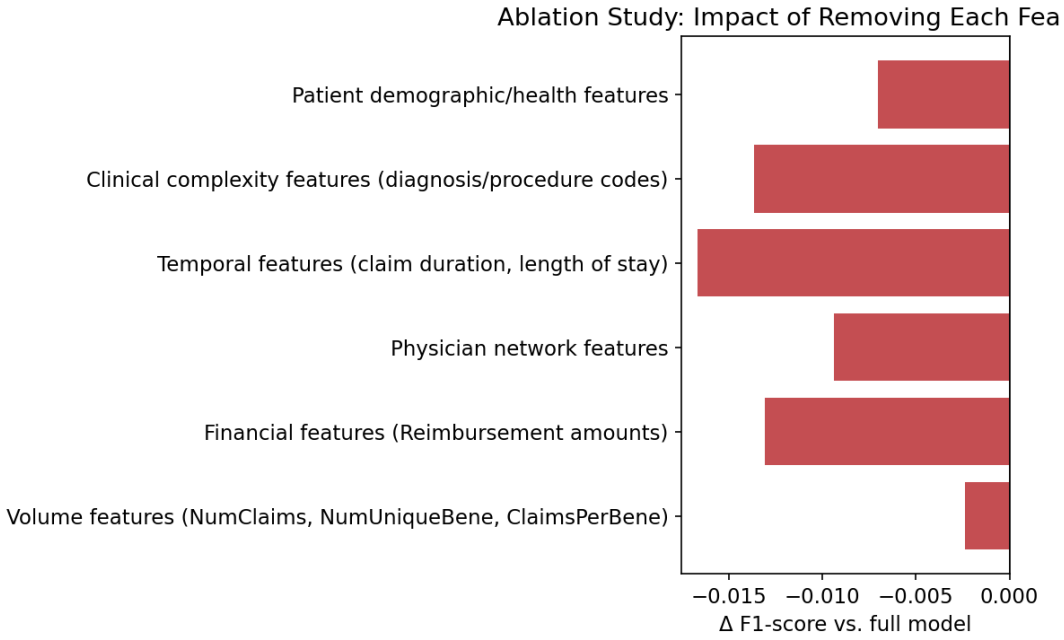

Figure 2. Ablation study: change in F1-score when each feature domain is removed from the full model.

Under synthetic Gaussian feature noise and simulated missing data — conditions directly relevant given that CMS attributes 77.17% of FY2025 Medicaid improper payments to insufficient documentation rather than confirmed fraud — XGBoost degraded gracefully and monotonically rather than collapsing sharply: ROC-AUC fell from 0.9506 with no noise to 0.8808 under the most extreme 50% noise condition tested, and to 0.9120 with 30% of feature values missing and mean-imputed. No catastrophic failure threshold was observed under either stressor, a practically reassuring finding for real-world deployment against imperfect billing or EVV documentation.

4.3 Explainability and Error Analysis

Because flagging a provider carries direct regulatory, financial, and reputational consequences, model decisions must be auditable by a human reviewer rather than presented as an opaque score (Rudin, 2019). SHAP TreeExplainer, which computes exact Shapley-value attributions for tree ensembles, identified total reimbursement as the single dominant predictive signal, followed by average length of stay, claims-per-beneficiary, and percent of claims that were inpatient. This is directly interpretable to a program-integrity analyst: providers who bill unusually large total amounts relative to their beneficiary count and typical stay duration are flagged — exactly the pattern a human auditor would manually search for, computed automatically and consistently across all 5,410 providers.

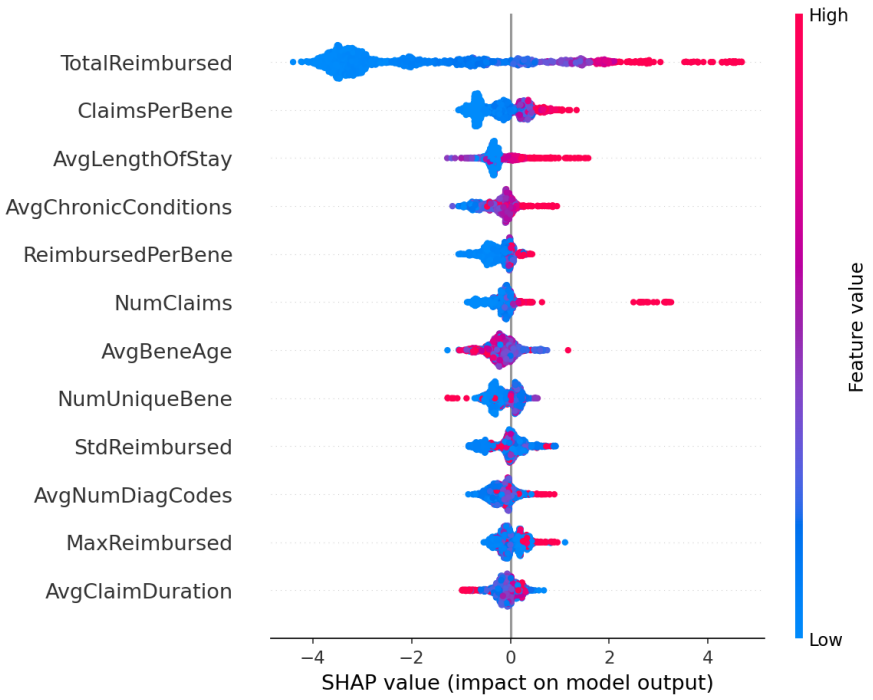

Figure 3. SHAP summary plot (top predictive features); red indicates a high feature value, blue a low one.

Of 127 confirmed-fraud providers in the held-out test set, Random Forest correctly identified 99 and missed 28, while flagging 79 legitimate providers as potential fraud. Missed cases had substantially lower mean total reimbursement ($124,149.64) than correctly identified cases ($811,861.92), and roughly one-third the claim volume and beneficiary count. This is an honestly reported and practically important limitation: the model under-detects lower-volume fraud relative to high-volume fraud, a plausible consequence of total reimbursement being the dominant predictive feature. It directly motivates incorporating additional, more granular features — such as EVV-derived per-visit timing anomalies — that do not require high aggregate volume to be individually suspicious.

5. Discussion

Three results converge on a single, actionable conclusion. First, temporal features were the most important single feature domain in ablation, yet were represented here only as coarse provider-level averages. Second, SHAP identified reimbursement magnitude and stay duration as dominant signals, both of which EVV can measure at far finer resolution. Third, structured error analysis showed that the model's principal blind spot — low-volume fraud — is exactly the regime where aggregate financial magnitude is least informative and where granular, per-visit behavioral signals would be expected to help most. Each of these findings independently points toward the same extension: replacing provider-level temporal averages with visit-level EVV timing data.

This alignment is what distinguishes the present contribution from a conventional benchmark study. The framework is deliberately compact (17 features), interpretable (exact Shapley attributions), statistically disciplined (paired testing and multi-seed stability rather than single-split claims), and robust (quantified, graceful degradation under documentation loss). These properties are not incidental; they are the properties a program-integrity system must have to be trusted in the regulated environment created by the EVV and prior-authorization mandates described in Section 1. A model that is accurate but unexplainable, or accurate but brittle under missing documentation, would fail precisely where CMS reporting says the real losses occur. Framed at the level of national impact, an explainable and deployment-ready system of this kind is a direct instrument for protecting public healthcare spending and strengthening the integrity of a multi-trillion-dollar sector of the U.S. economy.

Several limitations temper these conclusions and are stated plainly. The fraud labels are derived from historical exclusions and are therefore lagged and incomplete, so measured performance is best interpreted as discrimination against known exclusion

patterns rather than against all fraud. The dataset, while real and publicly available, mirrors Medicare claims conventions and may not capture every idiosyncrasy of state Medicaid billing. The physician-network features capture only first-order statistics, leaving richer graph structure unexploited. And the under-detection of low-volume fraud, though quantified honestly here, remains an open problem this framework does not yet solve.

6. Conclusion

This paper presented a real-data, statistically validated, explainable, and robustness-tested machine learning framework for Medicaid and Medicare provider fraud detection, directly motivated by CMS's finding that Medicaid improper payments reached $37.39 billion in fiscal year 2025, with the large majority attributable to documentation gaps that EVV systems are federally mandated to close. Using real, publicly available claims data spanning 5,410 providers, the study demonstrated that a compact, interpretable seventeen-feature provider-level representation supports strong, statistically stable fraud discrimination with sub-millisecond CPU inference latency; identified temporal, visit-timing features as the most predictive feature domain through systematic ablation; and honestly characterized the framework's current limitations, most notably its reduced sensitivity to lower-volume fraud.

Taken together, these findings provide a concrete, evidence-based, and appropriately cautious foundation for extending this methodology to real Electronic Visit Verification data, directly addressing a quantified and currently worsening federal program-integrity problem of national economic significance. Future research should pursue institutional-review-board-approved access to de-identified, visit-level EVV data to test directly whether granular timing features improve on the provider-level results reported here; extend the physician-network features into full graph-based representations; and conduct formal fairness and disparate-impact auditing prior to any real-world deployment.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Acknowledgments: Portions of this manuscript's text, code, and figure generation were drafted with the assistance of an AI language model. All data analysis was performed on real, verifiable, publicly available data; the authors reviewed, verified, and take full responsibility for all content, analysis, results, and conclusions presented. No experimental results, citations, or data were fabricated.

ORCID iD: Md Sibbir Hossain — 0009-0002-0795-4512; A T M Fokrule Hasan — [ORCID, if available]

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers.

References

1. Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: A survey. Data Mining and Knowledge Discovery, 29(3), 626–688. https://doi.org/10.1007/s10618-014-0365-y

2. Bauder, R. A., & Khoshgoftaar, T. M. (2017). Medicare fraud detection using machine learning methods. In 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 858–865). IEEE. https://doi.org/10.1109/ICMLA.2017.00-48

3. Bauder, R. A., & Khoshgoftaar, T. M. (2018). Medicare fraud detection using random forest with class imbalanced big data. In 2018 IEEE International Conference on Information Reuse and Integration (IRI) (pp. 80–87). IEEE. https://doi.org/10.1109/IRI.2018.00019

4. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

5. Centers for Medicare & Medicaid Services. (2021). Electronic Visit Verification (EVV). U.S. Department of Health and Human Services. https://www.medicaid.gov/medicaid/home-community-based-services/guidance/electronic-visit-verification-evv

6. Centers for Medicare & Medicaid Services. (2024). CMS Interoperability and Prior Authorization Final Rule (CMS-0057-F). https://www.cms.gov/newsroom/fact-sheets/cms-interoperability-prior-authorization-final-rule-cms-0057-f

7. Centers for Medicare & Medicaid Services. (2026). Fiscal year 2025 improper payments fact sheet. https://www.cms.gov/newsroom/fact-sheets/fiscal-year-2025-improper-payments-fact-sheet †

8. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

9. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

10. Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning (pp. 233–240). ACM. https://doi.org/10.1145/1143844.1143874

11. Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7, 1–30.

12. Dou, Y., Liu, Z., Sun, L., Deng, Y., Peng, H., & Yu, P. S. (2020). Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In Proceedings of the 29th ACM International Conference on Information and Knowledge Management (pp. 315–324). ACM. https://doi.org/10.1145/3340531.3411903

13. Elkan, C. (2001). The foundations of cost-sensitive learning. In Proceedings of the 17th International Joint Conference on Artificial Intelligence (pp. 973–978). Morgan Kaufmann.

14. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? Advances in Neural Information Processing Systems, 35, 507–520. https://arxiv.org/abs/2207.08815

15. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

16. Herland, M., Khoshgoftaar, T. M., & Bauder, R. A. (2018). Big data fraud detection using multiple Medicare data sources. Journal of Big Data, 5(1), 29. https://doi.org/10.1186/s40537-018-0138-3

17. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.

18. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In 2008 Eighth IEEE International Conference on Data Mining (pp. 413–422). IEEE. https://doi.org/10.1109/ICDM.2008.17

19. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data, 6(1), 1–39. https://doi.org/10.1145/2133360.2133363

20. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://arxiv.org/abs/1705.07874

21. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

22. Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). https://christophm.github.io/interpretable-ml-book/

23. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

24. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31, 6638–6648.

25. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

26. rohitrox. (2019). Healthcare provider fraud detection analysis [Data set]. Kaggle. https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis

27. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

28. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

29. Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. Information Fusion, 81, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

30. Van Belle, R., Baesens, B., & De Weerdt, J. (2025). Inductive inference of gradient-boosted decision trees on graphs for insurance fraud detection. arXiv. https://arxiv.org/abs/2510.05676 †

31. Wang, D., Lin, J., Cui, P., Jia, Q., Wang, Z., Fang, Y., Yu, Q., Zhou, J., Yang, S., & Qi, Y. (2019). A semi-supervised graph attentive network for financial fraud detection. In 2019 IEEE International Conference on Data Mining (ICDM) (pp. 598–607). IEEE. https://doi.org/10.1109/ICDM.2019.00070

32. Zafari, B., & Sundaram, D. (2022). Unsupervised machine learning for explainable health care fraud detection. arXiv. https://arxiv.org/abs/2211.02927 †