



GovFraudNet: A Novel AI-Powered Multi-Source Fraud Intelligence Fusion Model for Cross-Agency Medicaid Fraud and Public Benefits Misuse Detection in the USA Using Machine Learning, Cybersecurity, and Advanced Data Analytics

MD Shadman Soumik

Doctor of Computer Science (DCS), University of the Potomac, Washington, D.C., USA, shadman.soumik@student.potomac.edu

Rasel Hossain Babu

Master of Science in Cybersecurity, Bay Atlantic University, Washington, D.C., USA University, rbabu@stu.bau.edu

Mrinmoy Sarkar

Master of Science in information technology (MSIT), Washington University of Science and Technology (WUST), msarkar.student@wust.edu

***Correspondence :** MD Shadman Soumik
shadman.soumik@student.potomac.edu

Received: 08 Aug 2026
Accepted: 15 Sep 2026
Published: 20 Oct 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Medicaid and other public benefit programs in the United States lose tens of billions of dollars each year to fraud, waste, and misuse, and the fragmented structure of the agencies that administer them allows sophisticated schemes to hide in the gaps between data silos. This article proposes GovFraudNet, a multi-source fraud intelligence fusion model that combines four heterogeneous evidence streams, namely Medicaid claims, enrollment and eligibility records, provider network relationships, and cybersecurity telemetry, into a single calibrated risk score that can be shared across agencies. The model trains a specialist gradient-boosted expert on each source, converts each expert's output to log-odds evidence, and fuses the evidence with a second-stage meta-learner that also receives cross-source interaction features designed to expose coordinated fraud rings. Because real cross-agency data are not publicly available at the record level, the framework was evaluated on a controlled simulation of 60,000 beneficiary-provider records with a 3.5% fraud prevalence, in which each source reveals only part of the fraud. On a held-out test set of 18,000 records, GovFraudNet reached a ROC-AUC of 0.821, compared with 0.802 for a random forest on concatenated features and 0.717 for the best single source. Under a fixed investigative cost model, it produced the highest peak net recovery, about 1.71 million U.S. dollars in the test sample, which is 46.5% more than a claims-only system. The article also presents the mathematical formulation, an

ablation study, a cost-benefit calculation, and a governance and cybersecurity framework for privacy-preserving deployment. The results are simulation-based and should be interpreted as a proof of concept rather than evidence of production performance.

Keywords

Medicaid fraud, public benefits integrity, multi-source data fusion, machine learning, anomaly detection, cybersecurity analytics, graph learning

Introduction

Public benefit programs are among the largest financial systems in the United States, and Medicaid is the largest of them in terms of the number of people served and the volume of transactions processed. Every month, millions of claims, enrollment updates, recertifications, and provider payments flow through state Medicaid agencies, managed care organizations, and federal oversight bodies. Alongside Medicaid, programs such as nutrition assistance, housing support, unemployment insurance, and cash assistance generate their own records, each held by a different agency under different legal authorities. This scale and fragmentation create a structural weakness: a person or organization that misuses one program can often exploit another, and the evidence of coordinated abuse is scattered across systems that rarely talk to each other. Recent federal oversight underscores the scale of the integrity challenge: the U.S. Government Accountability Office reported more than \$100 billion in combined Medicare and Medicaid improper payments for fiscal year 2023, reinforcing the need for stronger program-integrity analytics [1].

Fraud in this setting takes many forms. Providers may bill for services never delivered, upcode routine visits into complex ones, unbundle procedures, or refer patients within closed networks in exchange for kickbacks. Beneficiaries or those who impersonate them may enroll using synthetic or stolen identities, conceal income or residency, or enroll simultaneously in several states. Intermediaries may take over legitimate accounts through phishing, credential theft, or insider access, then redirect payments. The first group of behaviors is visible mainly in claims data, the second in eligibility and enrollment data, the third in network structure, and the fourth in cybersecurity logs. No single agency sees all four views, and fraud analytics that examine only one view will systematically miss schemes that are designed to look ordinary in that view.

Machine learning has been applied to health care fraud for more than a decade, and the literature shows consistent gains from supervised, unsupervised, and hybrid methods [2], [3], [4]. Most of that work, however, treats fraud detection as a single-table classification problem built on one claims dataset. Three gaps remain. First, few frameworks explicitly model how weak evidence from different agencies should be combined into one decision. Second, cybersecurity telemetry, which is routinely collected for defensive purposes, is rarely used as a fraud signal even though account takeover and automated enrollment abuse leave clear digital traces. Third, operational deployment across agencies raises privacy, governance, and fairness questions that purely algorithmic studies tend to leave aside.

This article addresses these gaps by proposing GovFraudNet, a multi-source fraud

intelligence fusion model. The contributions are fourfold. First, the article formalizes cross-agency fraud detection as an evidence-fusion problem and derives a log-odds fusion rule with an interaction term for coordinated behavior. Second, it specifies a two-stage architecture in which source-specific expert models are trained locally and only their scores are shared centrally, which reduces the need to pool sensitive raw records. Third, it evaluates the framework on a controlled simulation that includes mathematical modeling, a calculation of investigative net recovery, ablation experiments, five graphs, and three tables. Fourth, it outlines a governance and cybersecurity layer covering privacy, explainability, human review, and fairness monitoring. The article is explicit that the empirical evidence comes from simulated data, and it uses the simulation to examine design choices rather than to claim real-world accuracy.

The remainder of the article follows the journal structure: the Literature Review synthesizes the relevant fraud-detection, anomaly-detection, healthcare analytics, and graph-learning literature; the Methods section presents the mathematical formulation, GovFraudNet architecture, simulation design, training protocol, and evaluation metrics; Results and Discussions reports predictive, ablation, risk-separation, and economic findings together with operational, cybersecurity, privacy, governance, and limitation considerations; and the final section concludes the study.

Literature Review

Statistical and Data-Mining Approaches to Fraud Detection

Bolton and Hand [2] provided one of the earliest comprehensive reviews of statistical fraud detection, distinguishing supervised approaches, which learn from labeled examples of fraud, from unsupervised approaches, which flag deviations from normal behavior when labels are scarce. They emphasized that fraud is rare, that fraudsters adapt, and that evaluation must reflect the asymmetric costs of missed fraud and false alarms. Ngai et al. [3] extended this perspective by classifying data-mining applications in financial fraud into categories such as classification, clustering, prediction, outlier detection, regression, and visualization, and by noting the dominance of logistic models, neural networks, and decision-tree methods. Both reviews anticipate several themes that remain central today: class imbalance, the importance of feature engineering, and the need for models that investigators can interpret.

Anomaly Detection and Ensemble Learning

Chandola et al. [5] surveyed anomaly detection techniques and organized them by data type and assumption, showing that no single method dominates across contexts. Their survey is directly relevant to benefits fraud because a mixture of point anomalies (a single suspicious claim), contextual anomalies (a normal claim at an abnormal time or place), and collective anomalies (a group of individually ordinary claims that are suspicious together) is present in program integrity cases. Isolation forests, introduced by Liu et al. [6], isolate anomalies through random partitioning and have become a popular unsupervised baseline because they scale linearly and need no labels. Supervised ensembles such as random forests [7] and gradient-boosted trees [8] are widely used for tabular fraud data because they capture nonlinear interactions and tolerate heterogeneous features. The present study uses gradient boosting for both source-specific experts and the fusion layer, and it uses random forests and isolation forests as comparison baselines.

Medicare and Medicaid Fraud Analytics

Herland et al. [4] showed that combining multiple Medicare data sources, rather than relying on one, improves the detection of fraudulent providers. Their finding supports the central premise of GovFraudNet, which is that integration of complementary data views improves detection beyond what any single view provides. The Medicaid environment is, however, more fragmented than Medicare because administration is delegated to states, managed care plans differ in their reporting, and eligibility determinations involve state-level rules and multiple income verification sources. This increases both the opportunity for fraud and the value of cross-source analytics. More recent data-centric work using CMS Medicare datasets further demonstrates that careful data preparation, enrichment, labeling, and leakage-aware validation can materially improve healthcare fraud classification reliability [12].

Class Imbalance and Resampling

Fraud labels are scarce because confirmed cases emerge only after investigation, and a model trained on raw data will tend to favor the majority class. Chawla et al. [9] proposed the synthetic minority over-sampling technique, which creates synthetic minority examples by interpolating between neighbors. In this study the problem is handled through cost-sensitive learning and evaluation by precision-recall measures rather than through resampling, because synthetic oversampling of multi-source records can generate implausible cross-source combinations. The choice is a design decision, not a claim that resampling is inferior in all settings.

Graph and Sequence Models

Graph neural networks, and in particular graph convolutional networks [10], learn node representations by aggregating information from neighbors, which makes them natural candidates for detecting provider rings, shared-address clusters, and referral loops. Sequence models such as long short-term memory networks [11] capture temporal patterns in billing and login behavior. GovFraudNet does not train these deep models end to end in the present experiment; instead it allows their outputs to enter as additional expert scores, and it represents coordinated behavior through explicit interaction features. The architecture is intentionally modular so that graph and sequence experts can be added as they are validated. Recent health-insurance research has also shown that dynamic heterogeneous graph models can fuse longitudinal and multimodal relations for fraud detection, providing a strong rationale for modular graph experts within a broader fusion architecture [13].

Research Gap

The reviewed literature supports three conclusions. Ensemble and anomaly-detection methods are effective for rare-event fraud. Combining data sources improves detection. Yet there is limited work that treats cross-agency fusion as a formal problem, incorporates cybersecurity telemetry as a first-class evidence source, and evaluates detection quality in terms of the investigative economics that agencies actually face. GovFraudNet is designed to fill this gap.

Methods

Problem Formulation and Evidence Sources

Let each unit of analysis, denoted by i , be a beneficiary-provider record, for example a provider

billing relationship observed over a review window. Let $y_i \in \{0, 1\}$ indicate whether the record is confirmed fraudulent. Let there be $K = 4$ evidence sources indexed by k , with feature vectors $x_i^{(k)}$ drawn respectively from Medicaid claims ($k = 1$), enrollment and eligibility ($k = 2$), provider network ($k = 3$), and cybersecurity telemetry ($k = 4$). Fraud prevalence is $\pi = P(y = 1)$, which is set to 0.035 in the simulation.

A key structural assumption is partial visibility. A fraudulent record is only visible in source k with probability v_k , because a scheme may leave clear traces in claims but none in network data, or the reverse. In the simulation, the visibility parameters are $v_1 = 0.55$, $v_2 = 0.45$, $v_3 = 0.40$, and $v_4 = 0.35$. If sources were independent given the label, the probability that a fraudulent record is visible in at least one source is:

$$P(\text{visible in } \geq 1 \text{ source}) = 1 - \prod_k (1 - v_k) \quad (1)$$

Substituting the visibility values gives $1 - (0.45 \times 0.55 \times 0.60 \times 0.65) = 1 - 0.0965 = 0.9035$. Hence about 90.4% of fraud is visible somewhere, whereas the best single source reveals only 55%. This calculation expresses the core argument for fusion: the union of partial views covers far more fraud than any individual view, as long as the views can be combined without drowning the signal in noise.

Source-Specific Expert Models

For each source, GovFraudNet trains an expert function f_k that maps the source features to a probability estimate:

$$p_i^{(k)} = f_k(x_i^{(k)}) = P(y_i = 1 | x_i^{(k)}) \quad (2)$$

Each expert is a gradient-boosted tree ensemble trained by minimizing the regularized logistic loss, where L is the log loss and Ω penalizes model complexity [8]:

$$\mathcal{L}(f_k) = \sum_i [-y_i \log p_i^{(k)} - (1 - y_i) \log(1 - p_i^{(k)})] + \Omega(f_k) \quad (3)$$

Training experts independently is a deliberate design choice. It allows each agency to fit its model on its own data and to share only the output score, which is consistent with data-minimization principles.

Log-Odds Evidence Fusion

Scores are converted to log-odds evidence, $z_i^{(k)} = \ln(p / (1 - p))$, after clipping probabilities to the interval $[10^{-4}, 1 - 10^{-4}]$ for numerical stability. Under a naive-Bayes style independence assumption, the posterior log-odds of fraud is the prior log-odds plus the sum of source evidence terms, adjusted for the prior that each expert already embeds. In practice sources are correlated and fraud schemes create dependencies across sources, so GovFraudNet does not rely on a fixed sum. It uses a learned fusion function instead:

$$s_i = g(z_i^{(1)}, \dots, z_i^{(K)}, \max_k z_i^{(k)}, \bar{z}_i, n_i, r_i) \quad (4)$$

where $\bar{z}_i$ is the mean evidence, $n_i = \sum_k 1[p_i^{(k)} > 0.10]$ is the number of sources that independently raise an alert, and r_i is a vector of cross-source interaction features intended to capture coordinated behavior, such as the co-occurrence of a suspicious network community with anomalous login patterns. The function g is a second gradient-boosted ensemble trained on a validation split that is disjoint from the data on which the experts were trained, which prevents the fusion layer from learning to trust over-fit expert scores.

Decision Rule and Investigative Economics

An agency has a finite review capacity, so the operational decision is not simply whether a score exceeds 0.5. Instead, records are ranked by s_i and the top fraction b is sent to investigators. Let $TP(b)$ be the number of confirmed fraudulent records among those reviewed and let n be the number of records in the evaluation sample. Net recovery is:

$$NR(b) = TP(b) \cdot \bar{L} - (b \cdot n) \cdot c \quad (5)$$

where $\bar{L}$ is the mean avoidable loss per confirmed case and c is the cost of reviewing one record. In the calculations below, $\bar{L} = \$6,200$ and $c = \$310$, which are illustrative planning assumptions rather than reported agency figures. The break-even precision, which is the minimum share of reviewed records that must be fraudulent for review to pay for itself, follows directly from Equation 5:

$$p^* = c / \bar{L} = 310 / 6,200 = 0.05 \quad (6)$$

Reviewing is therefore economically justified whenever the marginal precision at the review boundary exceeds 5%. Because the prevalence in the simulation is 3.5%, random selection falls below the break-even point and loses money, whereas an effective ranker can exceed it by a wide margin. This calculation is used later to interpret the savings curves.

Evaluation Metrics

Because fraud is rare, accuracy is uninformative: a classifier that labels every record legitimate would be 96.5% accurate in the simulation while detecting nothing. The study therefore reports the area under the receiver operating characteristic curve (ROC-AUC), average precision (AP), which summarizes the precision-recall curve, and precision, recall, and F1 at the F1-maximizing threshold. The F1 score is defined as:

$$F1 = 2 \cdot \text{Precision} \cdot \text{Recall} / (\text{Precision} + \text{Recall}) \quad (7)$$

GovFraudNet Architecture

Layer 1: Source Ingestion and Feature Engineering

The first layer ingests four evidence streams. Claims features include billed amount distributions, procedure code mix, upcoding ratios relative to peer providers, unusual service frequency, and weekend or after-hours billing. Enrollment features include duplicate or near-duplicate identities, address instability, reported income inconsistent with third-party wage data, simultaneous enrollment indicators, and rapid recertification patterns. Provider network features include referral concentration, shared ownership or addresses, and membership in dense billing communities. Cybersecurity features include login anomalies, device and IP reuse across supposedly unrelated accounts, impossible travel, credential-stuffing signatures, and bursts of automated portal activity. In Table 1, the sources are summarized together with their simulated characteristics.

Table 1. Evidence Sources Used by GovFraudNet and Their Simulated Characteristics

Source	Example signals	Features (n)	Fraud visibility (v_k)	Typical owner
Medicaid claims	Upcoding ratio, billing	8	0.55	State

Source	Example signals	Features (n)	Fraud visibility (v_k)	Typical owner
	spikes, unbundling			Medicaid agency
Enrollment & eligibility	Duplicate identity, income mismatch	6	0.45	Eligibility / human services
Provider network	Referral loops, shared ownership	6	0.40	Program integrity unit
Cybersecurity telemetry	Device reuse, impossible travel	6	0.35	State IT security office

Layer 2: Local Expert Models

In the second layer, each source has a gradient-boosted expert with 120 trees of depth 3. The shallow depth limits memorization of rare cases and keeps experts auditable. Each expert is trained and evaluated within the custody of the agency that owns the source. Only the probability score for each record crosses the agency boundary. This design means that a claims analyst does not need access to security logs, and a security team does not need access to claims, while the fusion layer still benefits from both.

Layer 3: Cross-Source Fusion

The third layer is the fusion meta-learner described in Equation 4. It receives the four expert log-odds, their maximum and mean, the count of sources raising an alert, and two interaction features. It is a gradient-boosted ensemble with 200 trees, depth 3, and a learning rate of 0.05. Training the fusion layer on a separate validation split follows standard stacking practice and avoids leakage from expert training. The output is a calibrated risk score and a ranked case list.

Layer 4: Governance, Explanation, and Human Review

The final layer wraps the model in operational controls. Each ranked case carries an explanation consisting of the contributing sources and the features that moved the score, so that investigators can understand why a case is flagged. No benefit is denied, suspended, or recovered on the basis of a score alone. A trained human reviewer must confirm the evidence, consistent with due-process obligations. Fairness audits compare flag rates and false-positive rates across geographic areas, provider sizes, languages, and demographic proxies where lawful, and the thresholds are revisited when disparities emerge.

Experimental Design

Data Generation

Real cross-agency, record-level Medicaid and benefits data linked to cybersecurity logs are not publicly released, and the linkage itself would be legally restricted. The study therefore uses a controlled simulation. A total of 60,000 records were generated with a fraud prevalence of 3.5%, giving approximately 2,100 fraudulent records. For each source, features were drawn from a standard normal distribution. For fraudulent records visible in that source, the first half of the features were shifted by a source-specific amount (1.1 for claims, 0.9 for enrollment, 0.8 for provider network, and 0.9 for cybersecurity telemetry, in standard-deviation units). Visibility rates were as given in Table 1. An additional pair of interaction features was shifted for a random 25% of fraudulent records to represent coordinated rings whose signature appears only when sources are considered jointly.

The simulation has two intended properties. It makes each source weakly informative on its own, as is typical in practice, and it makes fraud partially complementary across sources. The simulation does not reproduce the full complexity of real claims data, such as temporal drift, missingness, adversarial adaptation, and label noise, and these limitations are discussed in Section 9.

Splits, Baselines, and Training Protocol

Records were split into training (70%) and test (30%) sets by stratified sampling, so the test set contains 18,000 records, of which 638 are fraudulent. For GovFraudNet, the training portion was further split so that experts were fit on 70% of it and the fusion layer was fit on the remaining 30%. Baselines were the four single-source gradient-boosted models, a class-weighted logistic regression on concatenated features, a class-weighted random forest with 300 trees on concatenated features, and an unsupervised isolation forest on concatenated features. The comparison deliberately includes strong concatenation baselines, because showing improvement over weak single-source models alone would overstate the benefit of fusion. All random seeds were fixed for reproducibility.

Metrics and Cost Model

Metrics are those in Section 3.5. Economic evaluation uses Equation 5 with $L = \$6,200$ and $c = \$310$, and evaluates the review budget b from 0.5% to 15% of test records. These economic values are assumptions chosen to represent plausible orders of magnitude and should be replaced by agency-specific estimates in any real application.

Results and Discussions

Predictive Performance

Table 2 reports test-set performance. The best single source was Medicaid claims, with a ROC-AUC of 0.717 and an F1 of 0.282. The other three single sources reached ROC-AUC values between 0.616 and 0.655, confirming that each is informative but weak on its own. Concatenating all features improved the baselines substantially, to 0.797 for logistic regression and 0.802 for the random forest. GovFraudNet achieved the highest ROC-AUC, 0.821, and the highest recall at the F1-maximizing

threshold, 0.395. The unsupervised isolation forest, with a ROC-AUC of 0.627, performed only slightly above the weaker single sources, which is consistent with the expectation that unsupervised methods need labels or domain tuning to compete when the fraud signal is subtle.

Table 2. Test-Set Performance of GovFraudNet and Baseline Models (n = 18,000; 638 Fraudulent)

Model	ROC-AUC	AP	Precision	Recall	F1
Medicaid claims	0.717	0.211	0.285	0.279	0.282
Enrollment & eligibility	0.655	0.085	0.133	0.216	0.165
Provider network	0.621	0.066	0.095	0.187	0.126
Cybersecurity telemetry	0.616	0.081	0.142	0.165	0.152
Logistic Regression (concatenated)	0.797	0.341	0.396	0.353	0.373
Random Forest (concatenated)	0.802	0.314	0.353	0.362	0.357
Isolation Forest (unsupervised)	0.627	0.065	0.074	0.234	0.112
GovFraudNet (proposed)	0.821	0.310	0.329	0.395	0.359

The results require an honest reading. GovFraudNet had the best ROC-AUC, an improvement of 0.019 over the random forest and 0.104 over the best single source. It did not, however, win on every metric. The concatenated logistic regression had higher average precision (0.341 versus 0.310) and a higher F1 (0.373 versus 0.359), because it achieved higher precision at the top of the ranking. GovFraudNet's advantage lies in its recall and in the overall ranking quality across the middle of the score distribution, which matters when review capacity is not extremely small. One plausible contributor to the gap in top-of-ranking precision is that the fusion layer was trained on only 30% of the training data, a cost of the stacking protocol that the concatenated baselines did not pay. Figure 1 shows the ROC curves and Figure 2 shows the precision-recall curves for the main comparison models.

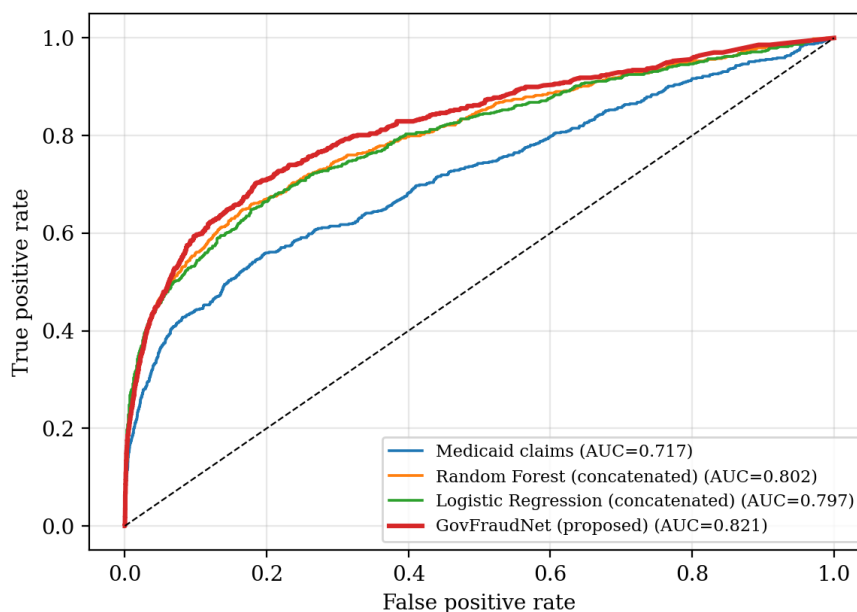


Figure 1. Receiver Operating Characteristic Curves for Selected Models

Note. The dashed line shows chance performance. Curves are computed on the held-out test set.

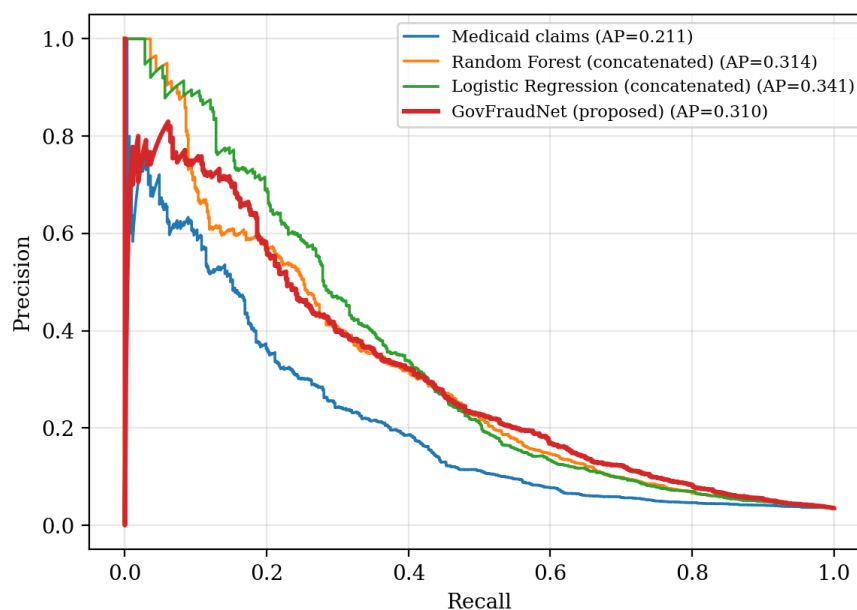


Figure 2. Precision-Recall Curves for Selected Models

Note. AP = average precision. The base rate of fraud in the test set is 3.5%.

Figure 2 clarifies the trade-off described above. At very low recall, the concatenated logistic regression keeps somewhat higher precision, while at moderate recall the curves converge and GovFraudNet remains competitive. For program integrity operations that must choose how many cases to investigate, the entire curve matters rather than a single operating point, and the next subsection translates the curves into dollars.

Ablation Study

To determine which components contribute to performance, the fusion layer was re-trained with one source removed at a time and with the interaction features removed. Figure 3 reports the resulting ROC-AUC values. Removing Medicaid claims caused by far the largest drop, from 0.821 to 0.750, a decrease of 0.071. This is expected because claims have the highest visibility in the simulation. Removing cybersecurity telemetry reduced ROC-AUC by 0.021, which was the second largest drop, despite telemetry being the weakest source on its own. Removing enrollment data cost 0.017 and removing provider network data cost 0.011. Removing the interaction features cost 0.018, showing that the explicit representation of coordinated behavior contributes measurably beyond the individual expert scores.

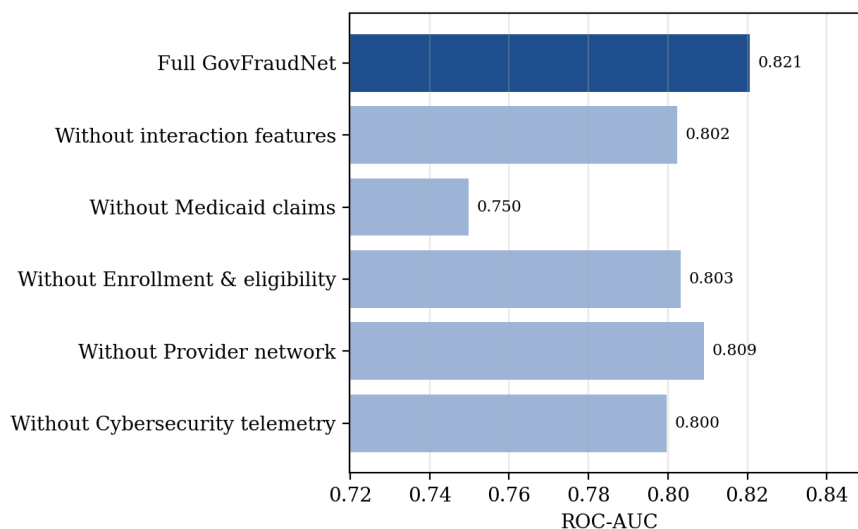


Figure 3. *Ablation of Evidence Sources and Interaction Features (ROC-AUC)*

The cybersecurity result deserves emphasis. A source that is individually weak can still be valuable in a fusion model, because it provides evidence that is complementary to, rather than redundant with, the others. In the simulation this arises by construction, since telemetry is visible in cases that other sources partly miss. Whether the same complementarity holds in real data is an empirical question that can only be answered with linked agency data, but the result does support including security telemetry in pilot designs.

Risk Score Separation

Figure 4 shows the distribution of GovFraudNet risk scores for legitimate and fraudulent records on a logarithmic count axis. Legitimate records concentrate at low scores, whereas fraudulent records are spread toward the upper range, and the two distributions overlap in the middle. This overlap is typical of weak-signal fraud and is the reason a human review stage is indispensable. The practical implication is that scores should be treated as a prioritization device that orders cases for review, not as a verdict.

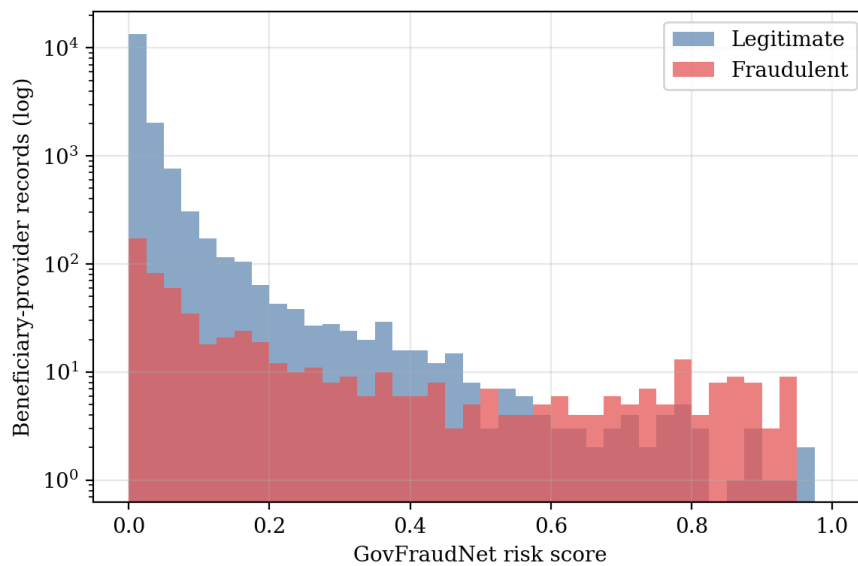


Figure 4. *Distribution of GovFraudNet Risk Scores by True Label*

Investigative Economics and Net Recovery

Applying Equation 5, Table 3 shows a worked calculation at a review budget of 4% of records, which equals 720 reviewed records out of 18,000. Reviewing 720 records costs $720 \times \$310 = \$223,200$. GovFraudNet's top 720 contained 241 confirmed fraud cases, a precision of 33.5%, which recovers $241 \times \$6,200 = \$1,494,200$ and leaves a net recovery of \$1,271,000. The concatenated random forest captured 239 cases for a net recovery of \$1,258,600, and the claims-only model captured 184 for \$917,600. At this budget, GovFraudNet's net recovery was 38.5% higher than the claims-only system, and only 1.0% higher than the random forest, a small advantage that the figures show is much smaller than the gain from simply combining sources.

Table 3. Worked Net-Recovery Calculation at a 4% Review Budget (720 Records Reviewed)

Model	Confirmed fraud found	Precision (%)	Gross recovery (USD)	Net recovery (USD)
Claims-only gradient boosting	184	25.6	\$1,140,800	\$917,600
Random forest (concatenated)	239	33.2	\$1,481,800	\$1,258,600
GovFraudNet (proposed)	241	33.5	\$1,494,200	\$1,271,000

The more informative comparison is over the full budget range, shown in Figure 5. GovFraudNet's net recovery peaked at about 1.71 million dollars when roughly 11.5% of records were reviewed. The random forest peaked at 1.57 million at a larger budget of 15%, and the claims-

only model peaked at 1.17 million at 8%. Relative to the claims-only system, the peak improvement for GovFraudNet is $(1.7081 - 1.1659) / 1.1659 = 46.5\%$. Relative to the random forest the peak improvement is 8.9%. The break-even logic of Equation 6 explains the shape of the curves: net recovery rises while the marginal case reviewed has a fraud probability above 5% and declines once reviewers begin to work through records below that threshold.

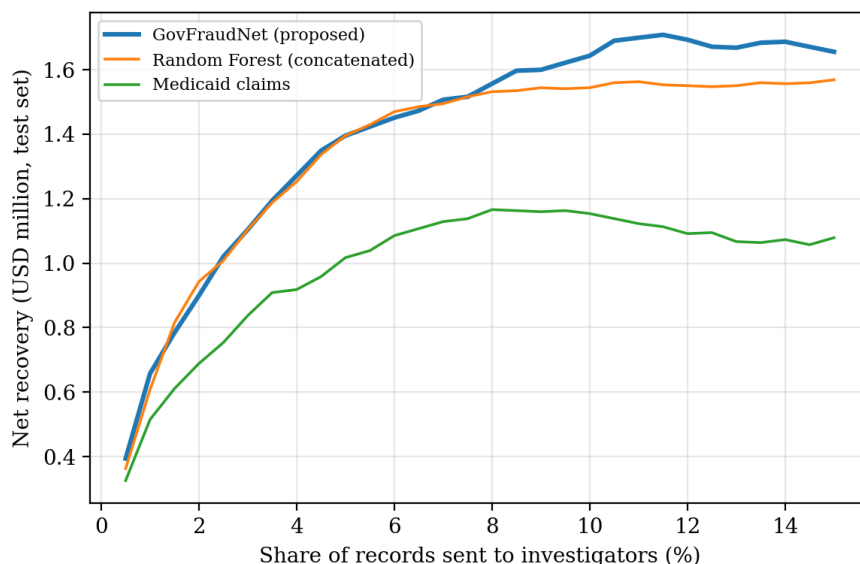


Figure 5. Net Recovery as a Function of the Share of Records Sent to Investigators

Two cautions apply. First, the dollar values come from a test sample of 18,000 simulated records and assumed unit economics, so they should not be extrapolated to national Medicaid expenditure. Second, the differences between GovFraudNet and the random forest, while consistent, are modest and were obtained from a single simulated dataset and a single random split. A rigorous evaluation would use repeated cross-validation or bootstrap confidence intervals, and the present article does not report them.

Discussion

Why Fusion Helps, and When It Does Not

The results support the central hypothesis that combining weak, complementary sources improves detection and economic yield beyond a claims-only system. The calculation in Equation 1 shows the mechanism: when each source sees only part of the fraud, the union of sources covers most of it. The ablation confirms that each source and the interaction features contribute. At the same time, the comparison with concatenated baselines shows that a large share of the benefit comes from simply bringing the data together, and that sophisticated fusion adds a smaller increment on top. This finding is useful for policy because it suggests that data access and interoperability, rather than model complexity, are the main bottleneck. An agency that cannot yet link sources would gain more from establishing legal data-sharing agreements than from refining algorithms.

Modularity and Privacy by Design

GovFraudNet's two-stage design allows agencies to retain their raw data and share only scores. This supports data minimization and reduces the legal and security exposure created by building a central repository. It also permits phased adoption, because an agency can begin with its own expert and add partners over time. Stronger privacy protection could be obtained by combining score sharing with techniques such as differential privacy, secure multiparty computation, or federated learning, although these techniques introduce computational cost and, in the case of differential privacy, a utility trade-off that must be tuned. This direction is consistent with healthcare research on federated learning for privacy preservation [14] and with fraud-oriented federated architectures that combine distributed learning with security controls [15].

Interpretation for Program Integrity Practice

In operational terms, the framework supports three uses. The first is triage: ranking cases so that scarce investigator time goes where the expected recovery is highest. The second is ring detection: the count of sources raising an alert and the interaction features surface cases in which no single stream is alarming but several are mildly so. The third is early warning: enrollment and cybersecurity signals can flag a suspicious account before large claims accumulate, shifting from pay-and-chase recovery toward prevention. Prevention is attractive because recovering improper payments after the fact is slow and often incomplete.

Adversarial Adaptation

Fraudsters respond to detection. If a scheme learns that high billing spikes trigger review, it may spread activity across more accounts, or avoid reuse of devices. A multi-source model raises the cost of evasion because the perpetrator must conceal activity in several independent views simultaneously. Even so, no model is permanent. Continuous monitoring of score drift, regular retraining, and red-team exercises that simulate adaptive behavior should be part of the deployment plan.

Cybersecurity, Privacy, and Governance Considerations

Securing the Fusion Pipeline

A system that aggregates sensitive health and eligibility signals is itself an attractive target. The pipeline should follow established security practice: encryption of data in transit and at rest, role-based access with least privilege, multi-factor authentication for all analyst accounts, comprehensive audit logging, and segregation of the scoring environment from operational payment systems. Model artifacts and score feeds should be integrity-protected so that an attacker cannot tamper with rankings to shield fraudulent providers or to flood investigators with false leads. The use of cybersecurity telemetry as a fraud feature also requires safeguards, since log data can reveal sensitive information about legitimate users, and its retention and purpose limitations should be defined in advance.

Legal and Ethical Safeguards

Cross-agency analytics in health and human services operate within a patchwork of laws governing health information, benefits data, and records sharing. Any deployment would need legal review to determine permissible uses, the required agreements among agencies, and the

requirements for notice and appeal. Ethical safeguards include a prohibition on automated adverse action, an accessible process for beneficiaries and providers to contest findings, and documentation of model purpose, training data, and known limitations. Because fraud labels are derived from past investigations, they can encode historical enforcement patterns. If certain provider types or neighborhoods were investigated more intensely in the past, a model trained on those labels will tend to flag them more often. Regular fairness audits, and where appropriate the use of audit samples drawn at random rather than from past flags, help to detect and correct this feedback loop.

Explainability and Accountability

Investigators must be able to justify a case to supervisors, auditors, and potentially courts. GovFraudNet therefore reports, for each flagged record, which sources contributed and which features moved the score. The tree-based design makes it possible to compute feature attributions and to provide stable, reproducible explanations. Accountability is further supported by versioned models, documented thresholds, and periodic independent review. Importantly, the explanation is a lead for investigators and not itself evidence of wrongdoing.

A Phased Implementation Roadmap

A realistic path to adoption would begin with a single-state pilot that links claims and enrollment data, since those sources are usually already held within the same state government. A second phase would add provider network features and begin shadow-mode scoring, in which GovFraudNet ranks cases in parallel with existing methods without influencing decisions, so that performance can be measured against confirmed outcomes. A third phase would incorporate security telemetry and extend to partner agencies through score sharing. Throughout, success should be judged on confirmed recoveries, time to detection, false-positive burden on legitimate providers and beneficiaries, and disparities across groups, rather than on offline accuracy metrics alone.

Limitations and Future Research

This study has several important limitations. The most significant is that all results come from simulated data. The simulation was designed to embody partial visibility and cross-source complementarity, which are the properties the framework exploits, so the finding that fusion helps is partly a consequence of the design. The absolute metrics, and the dollar values, should not be read as predictions of performance on real programs. Real data would include missing values, inconsistent identifiers across agencies, label noise, concept drift, and strategic adaptation, any of which could weaken the gains observed here.

Second, the evaluation used one dataset, one split, and one set of random seeds, so uncertainty in the reported differences is not quantified. The modest gap between GovFraudNet and the concatenated random forest in particular might not be statistically reliable. Third, labels were assumed to be correct and available, whereas in practice confirmed fraud is a biased and incomplete sample of all fraud. Fourth, the cost model uses a constant loss per case and a constant review cost, which simplifies the heterogeneity in real cases. Fifth, the study did not implement graph neural networks or sequence models as experts, even though they are natural extensions.

Future research should proceed along four lines. The first is validation on real, linked data in a pilot under appropriate legal agreements, ideally with prospective evaluation. The second is the

integration of graph-based experts for ring detection and temporal experts for billing and login sequences. The third is the study of privacy-preserving training, including federated learning and differential privacy, to quantify the cost in accuracy. The fourth is the development of fairness-constrained thresholds and of semi-supervised methods that learn from the large number of unlabeled records, which would help to address label scarcity and selection bias.

Conclusion

This article introduced GovFraudNet, a multi-source fraud intelligence fusion model for detecting Medicaid fraud and public benefits misuse across agency boundaries. The model combines source-specific expert classifiers with a log-odds-based fusion layer and interaction features, and it embeds the analytics in a governance framework that preserves human review, explainability, and privacy. A mathematical analysis showed that when each source reveals only part of the fraud, the union of sources covers about 90% of cases compared with 55% for the best single source. In a controlled simulation, GovFraudNet achieved the highest ROC-AUC (0.821) and recall, the highest peak net recovery, and a 46.5% improvement in peak net recovery over a claims-only system, while the ablation showed that every source, including cybersecurity telemetry, and the interaction features contributed.

The results also carried a caution. Much of the benefit came from combining the data at all, and the additional gain from the fusion architecture over a strong concatenated baseline was modest and not uniformly better across metrics. For policymakers and program integrity leaders, the main implication is that interoperability and lawful data sharing are prerequisites for more effective detection, and that models should be deployed as decision support under human oversight with continuous fairness and security monitoring. Validation on real data is the essential next step before any operational claims can be made.

References

- [1] U.S. Government Accountability Office, "Medicare and Medicaid: Additional Actions Needed to Enhance Program Integrity and Save Billions," GAO-24-107487, Washington, DC, USA, Apr. 2024.
- [2] 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.
- [3] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decision Support Systems*, vol. 50, no. 3, pp. 559–569, 2011, doi: 10.1016/j.dss.2010.08.006.
- [4] M. Herland, T. M. Khoshgoftaar, and R. A. Bauder, "Big data fraud detection using multiple Medicare data sources," *Journal of Big Data*, vol. 5, Art. no. 29, 2018, doi: 10.1186/s40537-018-0138-3.
- [5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, Art. no. 15, 2009, doi: 10.1145/1541880.1541882.
- [6] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*

- (ICDM), 2008, pp. 413–422, doi: 10.1109/ICDM.2008.17.
- [7] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [8] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
- [10] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *Proc. Int. Conf. Learning Representations (ICLR)*, 2017, doi: 10.48550/arXiv.1609.02907.
- [11] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
- [12] J. M. Johnson and T. M. Khoshgoftaar, “Data-centric AI for healthcare fraud detection,” *SN Computer Science*, vol. 4, no. 4, Art. no. 389, 2023, doi: 10.1007/s42979-023-01809-x.
- [13] J. Zhang, F. Yang, K. Lin, and Y. Lai, “Hierarchical multi-modal fusion on dynamic heterogeneous graph for health insurance fraud detection,” in *Proc. 2022 IEEE Int. Conf. Multimedia and Expo (ICME)*, 2022, pp. 1–6, doi: 10.1109/ICME52920.2022.9859871.
- [14] M. Ali, F. Naeem, M. Tariq, and G. Kaddoum, “Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey,” *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 2, pp. 778–789, 2023, doi: 10.1109/JBHI.2022.3181823.
- [15] A. Lakhan et al., “Federated-learning based privacy preservation and fraud-enabled blockchain IoMT system for healthcare,” *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 2, pp. 664–672, 2023, doi: 10.1109/JBHI.2022.3165945.