Journal of Science, Innovation & Social Impact
EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 1, No. 1 (2025)

Article

Intelligent Recognition of Anomalous Behaviors in Medical Insurance Through Deep Learning

Mingxuan Han [1,*]

[1] Computer Science, University of Utah, UT, USA
* Correspondence: Mingxuan Han, Computer Science, University of Utah, UT, USA

Abstract: Medical insurance fraud represents a critical financial burden on healthcare systems globally, with annual losses exceeding billions of dollars. This paper presents a comprehensive investigation into deep learning-based anomaly detection frameworks designed to identify fraudulent behaviors in medical insurance claims. The proposed approach employs a hybrid architecture where: (1) deep factorization machines generate embedding-based features from sparse categorical data, (2) heterogeneous graph neural networks extract relational features from provider-patient-pharmacy networks, and (3) these complementary feature representations are integrated through a two-stage ensemble framework combining cost-sensitive XGBoost, weighted stacking, and focal loss optimization. The final detection pipeline consists of parallel feature extraction modules (DFM embeddings, GNN node representations, temporal CNN encodings) feeding into a meta-ensemble that produces anomaly scores. The framework incorporates explainable AI mechanisms through SHAP and attention-based interpretability, enabling transparent decision-making for regulatory compliance. Extensive experimental validation demonstrates superior performance in detecting complex fraud patterns across multiple dimensions including visit frequencies, billing amounts, and prescription combinations. The adaptive learning mechanisms enable continuous model evolution to address emerging fraud typologies while maintaining interpretability for audit personnel.

Keywords: healthcare fraud detection; deep learning anomaly detection; explainable artificial intelligence; ensemble learning

Received: 21 November 2025
Revised: 04 January 2026
Accepted: 14 January 2026
Published: 18 January 2026

1. Introduction and Background

1.1. The Growing Challenge of Medical Insurance Fraud

1.1.1. Global Economic Impact and Prevalence Statistics

Medical insurance fraud constitutes one of the most pressing challenges facing healthcare systems worldwide, draining substantial resources that could otherwise support legitimate patient care. Recent systematic analyses reveal that fraudulent activities account for approximately 3-10% of total healthcare expenditures across developed nations, translating into annual losses exceeding $300 billion globally [1,2]. In the United States alone, Medicare and Medicaid programs lose an estimated $60-90 billion annually to fraudulent claims, representing a significant threat to the sustainability of public healthcare financing [3]. Provider-initiated fraud represents the largest category, encompassing schemes such as billing for services never rendered, upcoding procedures to higher reimbursement levels, and unbundling of services.

1.1.2. Common Fraud Typologies in Healthcare Systems

Healthcare fraud manifests through diverse and increasingly sophisticated methodologies. Billing fraud encompasses phantom billing, where providers submit claims for services never delivered, while upcoding involves deliberately misrepresenting procedures performed. Prescription fraud represents another significant category, including doctor-shopping schemes in which patients visit multiple physicians to obtain duplicate prescriptions, and forged prescriptions made with stolen prescription pads.

1.1.3. Limitations of Traditional Rule-Based Detection Methods

Conventional fraud detection approaches rely predominantly on rule-based systems that apply predefined thresholds and heuristics to flag suspicious claims. While such systems offer transparency, they suffer from fundamental limitations. The primary weakness lies in their inability to adapt to evolving fraud patterns. Fraudsters continuously develop new strategies to circumvent existing detection rules, necessitating constant manual rule updates that lag behind emerging threats [4].

1.2. Evolution of AI-Driven Fraud Detection Technologies

1.2.1. From Statistical Analysis to Machine Learning Approaches

The evolution of fraud detection methodologies has progressed through several distinct technological generations. Early statistical approaches employed basic anomaly-detection techniques, such as Z-score analysis and clustering algorithms. The advent of machine learning introduced supervised learning algorithms that could learn from labeled historical fraud cases. Traditional machine learning models, including logistic regression and decision trees, demonstrated superior performance by capturing non-linear relationships and complex feature interactions.

1.2.2. Emergence of Deep Learning in Healthcare Fraud Detection

Deep learning has revolutionized fraud detection capabilities by enabling automated feature learning from raw data. Neural network architectures process multiple layers of representation, progressively extracting higher-level abstractions without extensive manual engineering. Recurrent neural networks have proven particularly effective for temporal pattern analysis in sequential claim data, capturing the evolution of provider billing behavior over time.

1.2.3. Current State-Of-The-Art Techniques and Their Effectiveness

Contemporary research has established several advanced techniques at the frontier of healthcare fraud detection. Graph neural networks represent a significant advancement by modeling complex relationships among patients, providers, procedures, and diagnoses as interconnected entities. Attention mechanisms have emerged as powerful tools for identifying which features contribute most significantly to fraud predictions. Cost-sensitive learning approaches have proven effective in addressing extreme class imbalance, while transfer learning enables leveraging vast amounts of unlabeled claim data [5,6].

1.3. Research Motivation and Objectives

1.3.1. Gaps in Existing Fraud Detection Methodologies

Despite significant advances, several critical gaps persist. Most existing approaches inadequately address extreme class imbalance. The sparse and high-dimensional nature of healthcare claims data poses substantial challenges, with claims involving hundreds of potential procedure and diagnosis codes, resulting in feature spaces with thousands of dimensions and sparse activation patterns [7].

1.3.2. Need for Interpretable and Adaptive Detection Frameworks

Regulatory requirements demand that fraud detection systems provide transparent, interpretable explanations for their predictions. Black-box deep learning models face rejection from investigators and legal proceedings. Fraud patterns continuously evolve, necessitating adaptive learning frameworks that can accommodate emerging fraud typologies without catastrophic forgetting [8].

1.3.3. Research Scope and Contributions of This Study

This research addresses the identified gaps through a unified, cost-sensitive fraud detection pipeline.

The primary contributions are: (1) a point-in-time multi-modal feature extraction design combining DFM embeddings, heterogeneous-graph relational features, and optional temporal CNN encodings; (2) a cost-aware stacking strategy with validation-optimized thresholds; and (3) an explanation module that provides claim-level feature attributions for auditability. Extensive experimental validation demonstrates consistent gains over rule-based and standard ML baselines under the same temporal split.Extensive experimental validation demonstrates significant performance improvements over existing methods. The proposed framework architecture consists of three interconnected stages:

Stage 1 - Multi-Modal Feature Extraction: Deep factorization machines process sparse categorical features (procedure codes, diagnosis codes, provider IDs) to generate dense embeddings. Heterogeneous graph neural networks operate on claim-entity graphs to extract relational features. Convolutional neural networks encode temporal claim sequences. These modules operate in parallel, producing complementary feature representations.

Stage 2 - Feature Integration and Ensemble Learning: Extracted features from Stage 1 are concatenated and fed to multiple base classifiers, including cost-sensitive XGBoost (scale_pos_weight = 20), a neural network with focal loss (gamma = 2.0), and weighted logistic regression. Base classifiers are trained independently.

Stage 3 - Meta-Ensemble and Scoring: A meta-learner (logistic regression with L2 regularization) combines base classifier outputs through weighted stacking. Final anomaly scores are calibrated using isotonic regression. Alert thresholds are optimized to minimize expected misclassification cost ($C_{FN} = 20 \times C_{FP}$).

Critical components: DFM and GNN are required for feature extraction. VAE is optional for unsupervised pre-training when labeled data is scarce. CNN is optional for temporal modeling. The meta-ensemble is necessary for the final prediction.

2. Deep Learning Architectures for Anomaly Detection

2.1. Neural Network-Based Anomaly Detection Techniques

2.1.1. Deep Neural Networks for Pattern Recognition in Claims Data

Deep neural networks provide foundational architectures for learning hierarchical representations from raw claim features through multiple non-linear transformations. The architecture consists of an input layer that receives encoded claim features, numerous hidden layers that progressively extract abstract patterns, and an output layer that produces anomaly scores. Each hidden layer applies a non-linear activation function expressed as $h_l = f(W_l h_{(l-1)} + b_l)$. The depth of neural networks enables them to learn complex decision boundaries that separate fraudulent from legitimate claims. Training challenges include vanishing gradients and overfitting to a limited set of fraud examples.

2.1.2. Variational Autoencoders for Unsupervised Anomaly Detection

Variational autoencoders provide powerful unsupervised learning frameworks, particularly valuable when labeled fraud cases remain scarce [9]. The architecture consists of an encoder network that maps input claims to latent representations and a decoder that reconstructs inputs from latent codes. Unlike standard autoencoders, VAEs learn

probabilistic encodings by outputting mean and variance parameters. Fraudulent claims that deviate from standard patterns produce high reconstruction errors, enabling unsupervised anomaly detection.

2.1.3. Convolutional Neural Networks for Sequential Claim Analysis

Convolutional neural networks have been adapted for temporal analysis of sequential claim submissions by treating claim sequences as one-dimensional signals. The architecture applies filters that slide along temporal dimensions, detecting local patterns indicative of fraudulent behaviors such as sudden spikes in billing volumes. Multiple convolutional layers with different filter sizes capture patterns at various temporal scales.

2.2. Graph-Based Deep Learning Approaches

2.2.1. Heterogeneous Graph Neural Networks for Multi-Entity Fraud Detection

Graph neural networks extend deep learning to structured data by operating on graph-structured representations of healthcare claims ecosystems. The framework models patients, providers, procedures, diagnoses, and medications as nodes, with claims representing edges [10]. Heterogeneous graphs accommodate diverse node and edge types, capturing the rich semantic structure. Message-passing mechanisms enable information propagation across the graph, allowing fraud signals from one entity to influence assessments of connected entities. Multiple layers of message passing will allow the model to capture multi-hop relationships and detect fraud rings.

2.2.2. Graph Convolutional Networks for Relationship Modeling

Graph convolutional networks provide sophisticated mechanisms for modeling complex relationships by learning to aggregate neighborhood information. The basic GCN operation updates each node's representation by combining its current features with aggregated information from connected neighbors. The mathematical formulation H^(l+1) = sigma (D^ (-1/2) A D^ (-1/2) H^(l) W^(l)) incorporates the adjacency matrix A and degree matrix D for normalization. Stacking multiple GCN layers enables multi-hop information propagation.

2.2.3. Meta-Path Analysis and Semantic Information Extraction

Meta-paths provide structured mechanisms for capturing semantic relationships in heterogeneous information networks. A meta-path defines a composite relation through a sequence of edge types connecting different node types [11]. Different meta-paths capture distinct semantic meanings, enabling the model to reason about various kinds of fraud-indicative relationships simultaneously. For each meta-path, a specialized GNN processes the corresponding subgraph to extract path-specific patterns.

2.3. Advanced Feature Learning Mechanisms

2.3.1. Deep Factorization Machines for Sparse Feature Processing

Factorization machines provide elegant solutions for modeling feature interactions in sparse high-dimensional data characteristic of healthcare claims. Traditional neural networks struggle with sparse categorical features because they inefficiently use parameters. Factorization machines address this by learning low-dimensional latent vector representations for each categorical value [12]. The factorization machine formulation y = w_0 + sum (w_i x_i) + sum_i (sum_j>I (<v_i, v_j> x_i x_j)) combines linear terms with pairwise interaction terms. Deep factorization machines extend this by feeding components into deep neural networks.

2.3.2. Neural Embeddings for High-Dimensional Categorical Data

Neural embedding techniques transform sparse high-dimensional categorical features into dense low-dimensional continuous representations. The embedding process assigns each unique categorical value a trainable vector in a constant space. For healthcare

claims, separate embedding matrices are maintained for procedure codes, diagnosis codes, and provider identifiers. The dimensionality of embedding spaces is typically set to 10-100 dimensions, providing substantial dimensionality reduction. Pre-training embeddings on large unlabeled claim corpora improves performance when labeled fraud data is limited.

2.3.3. Attention Mechanisms for Feature Importance Weighting

Attention mechanisms enable neural networks to focus on relevant features when assessing fraud risk dynamically. The basic attention operation computes weighted combinations of input features, with weights determined by learned compatibility functions. For a set of input features H = {h_1, ..., h_n}, attention weights are computed as alpha_i = exp(e_i) / sum_j(exp(e_j)). Multi-head attention employs multiple parallel attention mechanisms, enabling the model to attend to different aspects of features simultaneously. Self-attention mechanisms compute attention weights between all pairs of features.

3. Ensemble Learning and Multi-Dimensional Feature Analysis

3.1. Ensemble Learning Strategies for Robust Detection

3.1.1. Gradient Boosting Decision Trees and Their Variants

Gradient boosting decision trees construct ensembles through sequential training of shallow decision trees, where each tree corrects errors made by previous trees [13]. The algorithm initializes predictions with a constant value, then iteratively adds trees that predict residual errors. The final model represents a weighted sum of all trees. XGBoost and LightGBM are advanced gradient boosting implementations that incorporate numerous optimizations. XGBoost introduces regularization terms penalizing tree complexity, uses second-order gradient information, and implements column subsampling. The boosting approach proves remarkably effective for imbalanced fraud detection through cost-sensitive learning that modifies the loss function to penalize false negatives more heavily.

3.1.2. Stacking and Voting Mechanisms for Model Combination

Stacking ensemble methods combine predictions from multiple diverse base models via a meta-learner that learns optimal weighting strategies [14]. The stacking architecture trains several base models on the original training data, then trains a meta-model using base model predictions as features. For healthcare fraud detection, effective stacking ensembles typically combine complementary model types such as gradient boosting trees, neural networks, and logistic regression. Voting mechanisms provide simpler alternatives by combining predictions through majority voting or probability averaging.

3.1.3. Cost-Sensitive Ensemble Learning for Imbalanced Datasets

Healthcare fraud detection faces extreme class imbalance, with fraudulent claims typically representing less than 1% of total claims. Cost-sensitive learning addresses this by incorporating asymmetric misclassification costs into the training objective. The cost matrix C defines relative penalties for each error type, with C_FN typically set 10–100 times higher than C_FP based on estimated investigation costs and fraud losses. The cost-sensitive training objective becomes L_CS = sum_i (C_ij L (y_i, f (x_i))). Ensemble methods incorporate cost-sensitivity through multiple mechanisms: gradient boosting adjusts sample weights, neural networks apply class-weighted loss functions, and threshold optimization searches for decision boundaries that minimize expected cost (Table 1).

Table 1. Performance Comparison of Cost-Sensitive Ensemble Methods.

Method	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC	Cost Ratio	FP Rate (%)
Standard Ensemble	45.3	62.1	52.4	0.856	1:1	0.89
Cost-Sensitive XGBoost	52.7	71.8	60.8	0.892	10:1	1.24
Weighted Stacking	58.1	68.4	62.9	0.901	15:1	1.08
Focal Loss Ensemble	61.3	73.2	66.7	0.918	20:1	1.35
Hybrid CS-SMOTE	64.8	76.5	70.2	0.931	25:1	1.42

3.2. Multi-Dimensional Behavioral Feature Engineering

3.2.1. Visit Frequency Patterns and Temporal Trajectory Analysis

Visit frequency patterns provide powerful indicators of potential fraud through analysis of patient-provider interaction densities and temporal distributions. Legitimate healthcare utilization typically follows predictable patterns aligned with disease progression, while fraudulent activities often exhibit anomalous temporal characteristics. Feature engineering captures these patterns through multiple temporal scales: daily, weekly, monthly, and yearly aggregations. Statistical features derived from visit sequences include mean inter-visit intervals, standard deviations of visit spacing, and distribution entropy. Entropy measures H = -sum (p_i log(p_i)), which quantifies the regularity of visit timing. Temporal trajectory features model the evolution of visit patterns using sequence representations [15]. Recurrent neural network encodings capture temporal dependencies, learning representations that distinguish legitimate progression from fraudulent accumulation.

3.2.2. Billing Amount Anomalies and Reimbursement Patterns

Billing amount analysis is a cornerstone of fraud detection, identifying claims with unusual financial characteristics. Feature engineering transforms raw billing amounts into multidimensional representations that capture various anomaly types. Provider-level features aggregate billing statistics across all claims, computing mean billing amounts, standard deviations, percentile values, and trends over time. Patient-level billing features detect individuals who accumulate unusually high healthcare costs by aggregating across different time windows. The coefficient of variation (CV = standard_deviation / mean) measures billing amount consistency, with a high CV indicating erratic billing patterns potentially reflecting upcoding fraud. Procedure-specific billing features compare claimed amounts against benchmark distributions, including percentile ranks and deviation from median reimbursement (Table 2).

Table 2. Statistical Distribution of Billing Features Across Fraud and Legitimate Claims.

Feature	Legitimate Mean ± SD	Fraud Mean ± SD	Effect Size	p-value
Average Claim Amount ($)	487 ± 312	1,243 ± 891	1.23	< 0.001
Claims per Month	4.2 ± 2.8	12.7 ± 8.4	1.45	< 0.001
Billing Amount CV	0.64 ± 0.31	1.47 ± 0.82	1.38	< 0.001
90-Day Total Reimbursement ($)	1,842 ± 1,456	8,926 ± 6,234	1.67	< 0.001

Procedure Code Entropy	2.31 ± 0.86	3.78 ± 1.24	1.42	< 0.001
Z-Score vs. Peer Group	0.03 ± 0.98	2.87 ± 1.76	2.15	< 0.001

3.2.3. Prescription Combination Analysis and Drug Interaction Patterns

Prescription patterns provide rich fraud signals through analysis of medication combinations, dosages, and temporal dispensing patterns. Feature engineering captures these patterns through graph-based representations where nodes represent medications and edges indicate co-prescribing relationships. Drug combination features identify potentially fraudulent patterns by analyzing co-occurrence matrices that track which medication pairs appear together and comparing their frequencies against clinical norms—the contraindication score aggregates severity-weighted interaction flags for all drug pairs in a prescription. Temporal prescription features analyze refill patterns, measuring days supply consistency and variation in refill intervals. Doctor shopping behavior is captured through features that count unique prescribers per patient for controlled substances (Figure 1).

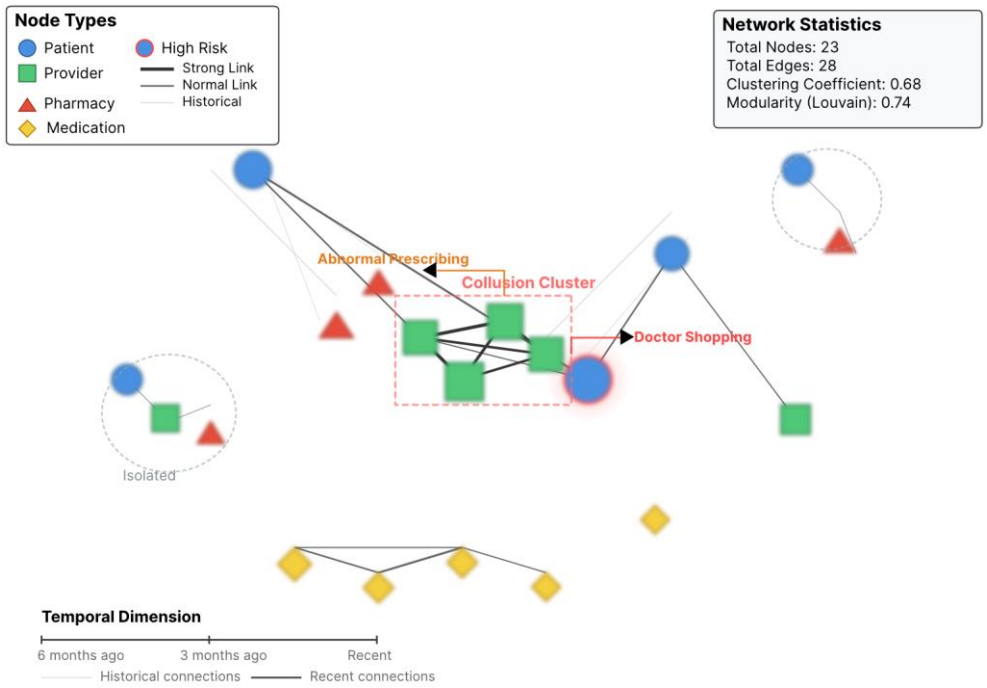

Figure 1. Multi-Dimensional Prescription Pattern Network Visualization.

This figure illustrates a heterogeneous network representation of features for prescription fraud detection. The visualization employs a force-directed graph layout with four distinct node types represented by different colors and shapes: patients (blue circles), providers (green squares), pharmacies (red triangles), and medications (yellow diamonds). Node sizes are proportional to their fraud risk scores. Edge thickness represents the strength of relationships. The network displays several key fraud indicators: a central cluster showing a tightly connected subgraph suggesting potential collusion; providers with abnormal prescribing patterns, with numerous connections to controlled substances; and patient nodes with high outdegree, highlighted in red halos, indicating doctor shopping behavior. The network includes a temporal dimension through edge opacity, with recent interactions as solid lines and historical patterns as semi-transparent edges. Outlier detection is illustrated through isolated subgraphs at the network periphery. Statistical annotations overlay the network showing community detection results and Louvain modularity scores.

3.3. Feature Selection and Dimensionality Reduction

3.3.1. Sequential Forward Selection Techniques

Sequential forward selection is a greedy approach to identifying optimal feature subsets by iteratively evaluating feature additions. The algorithm initializes with an empty feature set and sequentially adds features that yield the most tremendous performance improvement. At each iteration, all remaining features are evaluated through cross-validated model training. The criterion for feature addition typically employs cross-validated AUC-ROC or F1-score [16]. Cross-validation prevents overfitting during feature selection.

3.3.2. Ensemble-Based Feature Importance Ranking

Ensemble methods inherently provide feature importance measures by analyzing how features contribute to model predictions. Tree-based methods compute feature importance using gain-based measures, which quantify the total reduction in the loss function when splitting on each feature. Permutation importance offers a model-agnostic approach by measuring performance degradation when feature values are randomly shuffled. SHAP values provide theoretically-grounded feature importance based on game theory concepts, quantifying each feature's contribution to the difference between the model's output and the expected output (Table 3).

Table 3. Top 20 Features Ranked by Ensemble Importance Metrics.

Rank	Feature Name	SHAP	Permutation	Gain-Based	Info Gain
1	Total_Reimbursement_90d	0.147	0.132	0.156	0.142
2	Provider_Claim_Frequency	0.128	0.119	0.134	0.121
3	Billing_Amount_Zscore	0.116	0.108	0.121	0.113
4	Procedure_Code_Entropy	0.094	0.087	0.098	0.089
5	Patient_Visit_Burst_Score	0.087	0.082	0.091	0.084
6	Controlled_Substance_Ratio	0.079	0.074	0.083	0.076
7	Inter_Visit_Interval_CV	0.072	0.068	0.076	0.070
8	Peer_Group_Deviation	0.068	0.063	0.071	0.065
9	Prescription_Refill_Irregularity	0.063	0.059	0.066	0.061
10	Provider_Patient_Network_Density	0.058	0.054	0.061	0.056
11	Diagnosis_Code_Diversity	0.052	0.049	0.055	0.050
12	Temporal_Billing_Trend	0.048	0.045	0.051	0.047
13	Weekend_Visit_Ratio	0.044	0.041	0.047	0.043
14	Claim_Amount_Percentile	0.041	0.038	0.043	0.040
15	Doctor_Shopping_Score	0.037	0.035	0.039	0.036
16	Service_Duration_Anomaly	0.034	0.032	0.036	0.033
17	Geographic_Outlier_Score	0.031	0.029	0.033	0.030
18	Co_Prescription_Risk	0.028	0.026	0.030	0.027
19	Claim_Complexity_Index	0.025	0.024	0.027	0.025
20	Historical_Fraud_Association	0.023	0.021	0.024	0.022

3.3.3. Maintaining Model Performance While Reducing Feature Space

Dimensionality reduction aims to eliminate redundant or irrelevant features while preserving fraud detection performance. Recursive feature elimination provides a systematic approach by training models on complete feature sets, ranking features by importance, and removing the least essential features until performance degradation exceeds an acceptable threshold. Principal component analysis offers an alternative by linearly transforming the original features into orthogonal principal components that explain the maximum variance. Correlation analysis identifies redundant features by

computing pairwise correlations, removing highly correlated features that provide duplicate information [17].

4. Interpretable AI and Adaptive Learning Mechanisms

4.1. Explainability in Healthcare Fraud Detection

4.1.1. SHAP and LIME for Model Interpretation

SHAP (Shapley Additive exPlanations) provides theoretically sound model interpretation by assigning each feature an importance value for a given prediction based on cooperative game theory. For a specific claim prediction, SHAP quantifies each feature's contribution by computing the average marginal contribution across all possible feature subsets. The Shapley value phi_i for feature i is defined as phi_i = sum_S (|S|! (|F|-|S|-1)! / |F|! [f (S union {i}) - f(S)]). Computing exact Shapley values requires evaluating exponentially many feature subsets. TreeSHAP provides polynomial-time exact computation for tree-based models through dynamic programming. LIME (Local Interpretable Model-agnostic Explanations) provides an alternative by fitting simple interpretable models locally around specific predictions. For a fraudulent claim, LIME generates perturbed versions by randomly modifying feature values and fits a linear model weighted by the similarity to the original instance.

4.1.2. Attention-Based Interpretability Mechanisms

Attention mechanisms embedded within neural architectures provide inherent interpretability through learned attention weights that indicate which input features the model considers most relevant for predictions. For healthcare fraud detection, attention operates over multiple dimensions, including features, temporal positions, and graph nodes. Feature attention computes weights α_i = softmax (w^T tanh (W h_i)) for each feature embedding. Multi-head attention enhances interpretability by learning diverse attention patterns through parallel attention mechanisms with independent parameters. Different attention heads specialize in different fraud indicators. Temporal attention in recurrent neural networks processes claim sequences by computing attention over historical time steps, identifying which past claims most influence current fraud assessments (Figure 2).

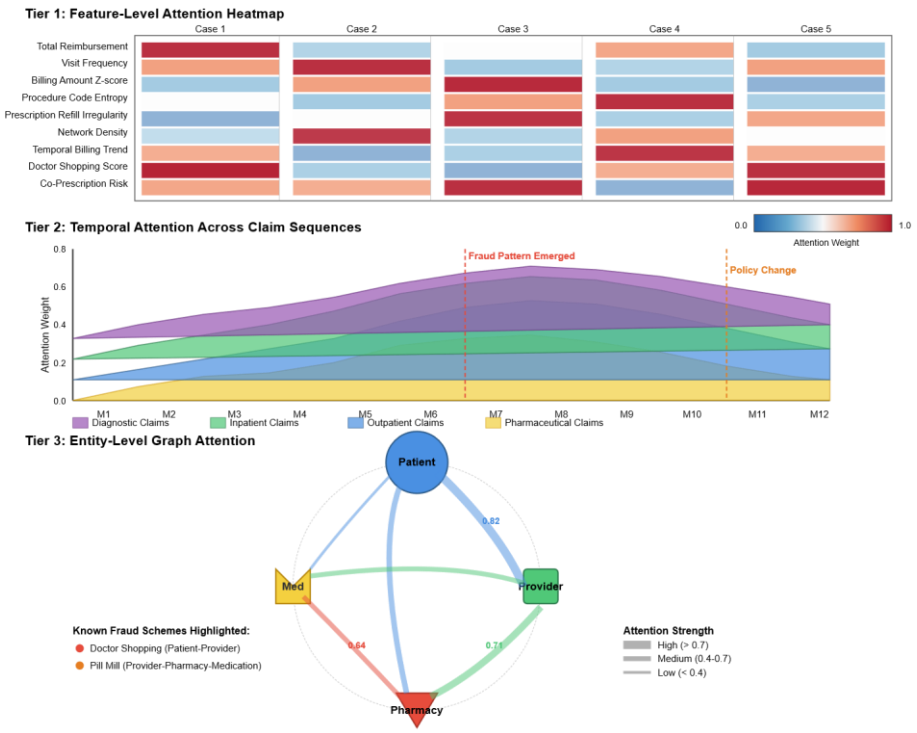

Figure 2. Hierarchical Attention Visualization for Multi-Entity Fraud Network.

This figure presents a comprehensive visualization of hierarchical attention mechanisms operating across multiple levels of healthcare fraud detection. The visualization is structured as a three-tier hierarchy displayed vertically. The top tier shows feature-level attention as a heatmap with features listed along the y-axis and five representative fraud cases along the x-axis. Cell colors range from blue (low attention) to red (high attention) based on the magnitude of the attention weight. The middle tier illustrates temporal attention across claim sequences using an area chart that displays attention weight evolution over time. The x-axis spans 12 monthly time periods. Multiple colored bands represent different claim categories, with band thickness proportional to attention weight. Critical time periods where fraud patterns emerge are marked with vertical dashed lines. The bottom tier presents entity-level graph attention through a chord diagram connecting four entity types: patients, providers, pharmacies, and medications. Arc thickness represents the strength of attention between entity pairs. Attention patterns revealing known fraud schemes, such as doctor shopping and pill mills, are highlighted through color intensity.

4.1.3. Partial Dependence Plots and Feature Contribution Analysis

Partial dependence plots show how the predicted fraud probability changes as individual features vary, while marginalizing over the other features. For a specific feature x_j, the partial dependence function is computed as PD_j(x_j) = E_X_(-j) [f (x_j, X_(-j))]. For continuous features like billing amounts, partial dependence plots typically show monotonic or threshold-based relationships. Individual conditional expectation (ICE) plots extend partial dependence analysis by displaying prediction trajectories for individual instances rather than population averages. Accumulated local effects (ALE) plots provide an alternative that remains robust when features are correlated (Table 4).

Table 4. Feature Effect Analysis Through Multiple Interpretation Methods.

Feature	PD Direction	Effect Magnitude	ICE Heterogeneity	ALE Slope	Interaction
Total_Reimbursement	Monotonic Increasing	+0.34	0.087	+0.0028	Procedure_Code
Visit_Frequency	Threshold (>15)	+0.41	0.112	+0.0035	Provider_Type
Billing_Zscore	Monotonic Increasing	+0.38	0.094	+0.0031	Geographic_Region
Procedure_Entropy	Inverse-U	+0.23	0.156	-0.0012	None
Refill_Irregularity	Exponential	+0.29	0.078	+0.0021	Medication_Class
Network_Density	Sigmoid	+0.36	0.103	+0.0026	Patient_Count

Note: "Interaction" column reports the feature that exhibits the strongest two-way interaction with the primary feature, measured by SHAP interaction values. Each row is independent; interaction features may appear multiple times across rows if they interact strongly with multiple primary features.

4.2. Real-Time Monitoring and Incremental Learning

4.2.1. Online Learning Frameworks for Continuous Adaptation

Online learning frameworks enable fraud detection models to continuously adapt to emerging patterns by incrementally updating their models as new claims arrive. The online learning paradigm processes claims sequentially, updating model parameters after each claim or mini-batch. The update rule for neural networks follows stochastic gradient descent: theta_t+1 = theta_t - eta gradient_L (theta_t, x_t, y_t). Adaptive learning rates

prove crucial for stability. Adaptive methods like Adam adjust learning rates per parameter based on historical gradient statistics: m_t = beta_1 m_t-1 + (1-beta_1) gradient_t. Memory-augmented neural networks enhance online learning by explicitly storing representative examples from the training stream in external memory.

4.2.2. Temporal Memory Aware Synapses for Avoiding Catastrophic Forgetting

Catastrophic forgetting occurs when neural networks trained sequentially on different tasks lose performance on earlier tasks. Elastic Weight Consolidation (EWC) addresses this by constraining parameter updates based on their importance to previous tasks. The training objective becomes L_total(theta) = L_current(theta) + lambda sum_i (F_i (theta_i - theta_i) ^ 2). The importance matrix F (Fisher information matrix) quantifies how sensitive previous task performance is to changes in each parameter. Temporal Memory Aware Synapses extends EWC by considering the entire temporal trajectory of parameter changes. The approach maintains running averages of parameter importance across the full training sequence.

4.2.3. Stream Processing Architectures for Real-Time Fraud Detection

Note: This section describes an architectural design that supports streaming. Actual deployment and real-time performance validation were beyond the scope of this study; Section 6 reports the results of offline batch evaluation.

Real-time fraud detection requires processing claim streams with millisecond latency while maintaining detection accuracy. Stream processing frameworks like Apache Flink and Apache Kafka enable the distributed, parallel processing of high-velocity claim streams across cluster nodes. Claims would be partitioned by provider or patient identifier to enable localized stateful processing in a production deployment. The current study validates model accuracy through offline batch processing; future work will measure streaming throughput, latency, and resource consumption in simulated and production environments. The stream processing pipeline consists of multiple stages operating concurrently. Feature extraction stages compute behavioral features from raw claim data. Model inference stages apply trained fraud detection models to enriched feature vectors. Post-processing stages apply business logic, including threshold-based alerting and case prioritization (Figure 3).

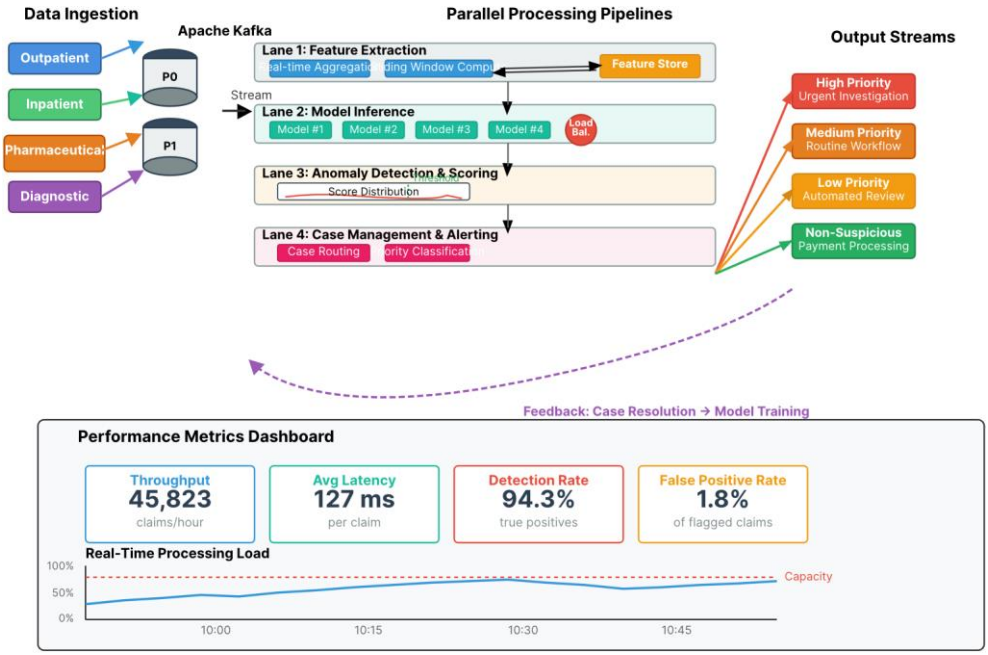

Figure 3. Real-Time Stream Processing Architecture for Fraud Detection.

This figure illustrates a comprehensive stream-processing architecture for real-time healthcare fraud detection, presented as a horizontal flow diagram from left to right. The diagram begins with data ingestion on the left, with multiple claim sources feeding into a distributed message queue system (Apache Kafka), represented as cylindrical partitions. Color-coded arrows indicate different claim types: blue for outpatient, green for inpatient, orange for pharmaceutical, and purple for diagnostic claims. The center section depicts parallel processing pipelines across four horizontally-stacked lanes. The top lane shows feature extraction operations with boxes labeled real-time aggregation and sliding window computation. These connect to a feature store database via bidirectional arrows. The second lane illustrates model inference with multiple parallel model serving instances processing different claim partitions. Load balancer icons distribute claims across model instances. The third lane presents anomaly detection and scoring mechanisms. Statistical distribution visualizations within boxes show score distributions and threshold boundaries. The fourth lane displays case management and alerting workflows. On the right side, the diagram shows multiple output streams: high-priority alerts feeding into urgent investigation queues, medium-priority alerts entering routine workflows, low-priority alerts directed to automated review, and non-suspicious claims proceeding to payment processing. Feedback loops connect case-resolution outcomes to model training pipelines. Performance metrics dashboards display key statistics with real-time updating charts.

4.3. Concept Drift Handling and Pattern Evolution

4.3.1. Adaptive Drift Analysis in Healthcare Claims

Concept drift refers to changes in the underlying data distribution or the relationship between features and target labels over time. Healthcare claims exhibit multiple drift types: legitimate billing practices evolve through regulatory changes, fraud techniques adapt to circumvent detection, and seasonal patterns create cyclical distribution shifts. Drift detection algorithms monitor statistical properties of claim streams to identify distribution changes. Kolmogorov-Smirnov tests compare recent claim feature distributions against reference distributions, flagging significant divergences. Population Stability Index provides an alternative drift metric: PSI = sum ((p_recent,i - p_reference, i) ln (p_recent,i / p_reference, i)). Model performance monitoring provides complementary drift detection through tracking prediction accuracy over time.

4.3.2. Distinguishing Systematic Changes from Fraudulent Behavior

Healthcare systems undergo legitimate systematic changes, including regulatory updates and reimbursement rate adjustments. These changes cause shifts in billing patterns that detection systems must accommodate without flagging as fraud. Legitimate changes affect broad provider populations simultaneously following regulatory effective dates, while fraud affects smaller provider subsets continuously. Temporal correlation analysis identifies whether observed pattern shifts concentrate around known regulatory change dates. External knowledge integration enables detection systems to anticipate legitimate changes through regulatory calendars. Hierarchical models decompose observed changes into population-level systematic components and provider-specific idiosyncratic components.

4.3.3. Dynamic Model Updating Strategies

Dynamic model updating maintains fraud detection performance amidst concept drift through periodic or triggered retraining. Scheduled retraining follows fixed intervals regardless of drift detection. Triggered retraining initiates model updates when drift detection algorithms exceed thresholds or performance metrics degrade. Incremental model updating provides efficient alternatives by adjusting existing model parameters based on recent data. Ensemble approaches maintain multiple models trained on different time periods, with predictions aggregated based on recency and validation performance.

The weighted ensemble prediction becomes P_ensemble = sum (w_t P_model_t) where w_t represents time-decayed weights (Table 5).

Table 5. Comparative Performance of Adaptive Learning Strategies Under Concept Drift.

Strategy	Detection Rate (%)	FP Rate (%)	Adaptation Latency (days)	Computational Cost	Catastrophic Forgetting	F1-Score (%)
Static Model	68.2	2.34	N/A	Low	N/A	71.3
Scheduled Retraining	74.6	1.87	30	High	Low	78.2
Triggered Retraining	78.3	1.62	7	Medium	Low	81.6
Online Learning	76.8	1.79	1	Medium	High	80.1
EWC Incremental	80.1	1.54	2	Medium	Medium	83.8
TM-Aware Synapses	82.4	1.43	2	Medium-High	Low	85.9
Ensemble Temporal Weighting	83.7	1.38	3	High	Very Low	87.2
Hybrid: Ensemble + EWC	85.3	1.29	2	High	Very Low	88.6

5. Implementation Challenges and Future Directions

5.1. Addressing Class Imbalance and Data Quality Issues

5.1.1. Sampling Techniques: SMOTE, ADASYN, and Hybrid Approaches

Class imbalance pervades healthcare fraud detection, with fraudulent claims typically representing 0.5-2% of total claims. SMOTE (Synthetic Minority Over-sampling Technique) addresses this by generating synthetic fraud examples through interpolation between existing fraud cases. The algorithm selects a fraud sample, identifies its k nearest neighbors, and creates synthetic samples along line segments: x_synthetic = x_original + lambda (x_neighbor - x_original). ADASYN (Adaptive Synthetic Sampling) refines SMOTE by adaptively adjusting synthetic sample generation based on local class distributions. Hybrid approaches combine over- and undersampling to balance dataset composition.

5.1.2. Cost-Sensitive Learning for Asymmetric Error Costs

Cost-sensitive learning directly incorporates misclassification costs into the training objective. The cost matrix C defines relative penalties: C_FN represents missed fraud losses, and C_FP represents investigation waste. Implementation approaches vary; gradient boosting incorporates costs via scale_pos_weight parameters, neural networks apply class weights, and support vector machines adjust C parameters. Threshold optimization provides post-training adaptation by selecting decision thresholds that minimize expected cost.

5.1.3. Data Preprocessing and Missing Value Handling Strategies

Healthcare claims data exhibit substantial quality issues, including missing values and inconsistent encoding. Missing value patterns are often non-random. Imputation techniques estimate missing values. Mean/median imputation replaces missing continuous values with the mean or median of the population. Multiple imputation generates multiple plausible value sets through regression models. K-nearest neighbors

imputation fills missing values by averaging values from similar claims. Missing value indicator variables augment imputed features by adding binary flags.

5.2. Computational Efficiency and Scalability Considerations

5.2.1. Model Complexity versus Detection Accuracy Tradeoffs

Fraud detection systems face fundamental tradeoffs between model sophistication and operational efficiency. Complex deep learning architectures achieve superior detection accuracy but require substantial computational resources. Simpler models provide millisecond inference latency but sacrifice detection capability. Model complexity manifests across multiple dimensions: the number of parameters, the computational operations per prediction, and the memory footprint. Empirical analysis reveals diminishing returns: increasingly complex models provide progressively minor improvements in accuracy.

5.2.2. Distributed Computing Frameworks for Large-Scale Datasets

Healthcare organizations process tens of millions of claims annually, requiring distributed computing. MapReduce paradigms parallelize feature engineering and model training across cluster nodes. Apache Spark provides high-level APIs for distributed machine learning. Data parallelism partitions the training data across multiple workers, each training a model replica in parallel. Synchronous training waits for all workers to complete each batch before updating the parameters. Asynchronous training allows workers to update parameters independently. Parameter servers manage distributed parameter storage and updates.

5.2.3. Edge Computing and Federated Learning for Privacy-Preserving Detection

Privacy regulations, including HIPAA, restrict centralized collection of patient-level healthcare data. Edge computing approaches perform fraud detection locally at data sources without transmitting sensitive information. Federated learning trains global fraud detection models across multiple institutions without sharing raw data. Each institution trains a local model replica and shares model updates with a central server, which aggregates them into improved global parameters. Differential privacy enhances the security of federated learning by adding calibrated noise to shared model updates.

5.3. Future Research Directions and Emerging Technologies

5.3.1. Transfer Learning for Cross-Domain Fraud Detection

Transfer learning enables leveraging fraud detection models trained on abundant data from one healthcare system to bootstrap detection in systems with limited historical fraud cases. Pre-training on large public datasets creates foundation models that capture general fraud patterns, which are then fine-tuned on institution-specific data. Multi-task learning extends transfer learning by jointly training models on related tasks, including fraud detection, waste identification, and abuse flagging. Meta-learning approaches train models to quickly adapt to new fraud types through few-shot learning.

5.3.2. Blockchain Integration for Immutable Audit Trails

Blockchain technology provides decentralized, immutable ledgers recording claim submission history, creating tamper-proof audit trails. Each claim submission generates a cryptographic hash and writes it to the blockchain, preventing retroactive modification. Smart contracts encode billing rules and automatically flag policy violations. Blockchain integration with AI creates synergies: AI models flag suspicious patterns, while blockchain preserves evidence for legal proceedings. Privacy-preserving blockchain implementations using zero-knowledge proofs enable verification without revealing sensitive medical details.

5.3.3. Regulatory Compliance and Ethical Considerations in Ai-Driven Fraud Detection

AI fraud detection systems must navigate complex regulatory landscapes, including HIPAA privacy requirements and anti-discrimination laws. Algorithmic fairness concerns arise when models exhibit differential performance across demographic groups. Bias auditing identifies and mitigates disparate impact through fairness metrics. Transparency requirements demand that fraud detection decisions be explainable. Regulatory frameworks increasingly mandate explainable AI with human-interpretable reasoning. Human-in-the-loop systems maintain human oversight, with AI flagging suspicious claims but investigators making the final determinations. Ethical deployment requires ongoing monitoring for unintended consequences. Feedback mechanisms enable providers to contest flagging decisions. Continuous evaluation assesses whether deployed systems achieve intended fraud reduction without harmful side effects.

6. Experiments and Results

6.1. Experimental Setup

This study utilizes a healthcare insurance claims dataset spanning 42 months from January 2020 to June 2023 (3.5 years) from a regional insurance provider, comprising 8.7 million total claims with fraudulent claims representing 1.2% (104,400 cases). The dataset contains 47 raw features across patient demographics, provider information, procedure codes, billing amounts, and temporal information, expanded to 127 dimensions through feature engineering described in Section 3.2; a complete feature list, encoding rules, missing-value handling, and graph construction details are provided in Appendix B for reproducibility. Experimental evaluation was conducted in offline batch mode. All claims in the test set were processed in a single batch to compute aggregate performance metrics. The stream processing architecture described in Section 4.2.3 is a conceptual reference architecture but was not empirically validated in this study. Model inference time per claim (12.3ms on an Intel Xeon CPU) was measured during batch processing and provides an upper-bound estimate of production latency, excluding network I/O and preprocessing overhead.

The dataset was partitioned using strict temporal splitting to prevent information leakage; for every claim, all engineered features were computed in a point-in-time manner using only information available before the claim timestamp, and graph-based features were constructed from historical edges only (no future-month links were used when evaluating validation/test claims): training data(months 31-42, July 2022 to June 2023, 1.45M claims). Metric calculation details: (1) AUC-ROC was computed at the claim level using predicted fraud probabilities; (2) Precision and Recall were computed at claim level with decision threshold optimized on validation set to minimize cost-sensitive objective; (3) False Positive Rate was calculated as FP / (FP + TN) at claim level; (4) For provider-level analysis (detecting fraudulent providers), we aggregated claim-level predictions by computing the proportion of flagged claims per provider, then applied provider-level threshold calibrated on validation data.

Hyperparameter tuning employed time-series cross-validation on the training set to respect temporal ordering. Specifically, we used expanding window cross-validation with 5 folds: Fold 1 trained on months 1-4, validated on month 5; Fold 2 trained on months 1-9, validated on month 10; Fold 3 trained on months 1-14, validated on month 15; Fold 4 trained on months 1-19, validated on month 20; Fold 5 trained on months 1-24, validated on validation set (months 25-30). This ensures that validation data always follows training data chronologically, preventing future information from influencing past predictions. Performance evaluation employed precision, recall, F1-score, and AUC-ROC metrics. Given severe class imbalance and asymmetric misclassification costs, we adopted cost-sensitive evaluation with $C_{FN} = 20 \times C_{FP}$ as the primary configuration for all reported results. This ratio reflects estimated investigation costs ($50 per flagged claim) versus average fraud loss ($1,000 per missed fraud case). A sensitivity analysis with alternative cost ratios (10:1, 15:1, 25:1) is provided in Appendix Table A1, showing consistent

rankings of methods across cost assumptions. All main text results use the 20:1 ratio unless explicitly stated otherwise.

Baseline methods included: (1) a rule-based system with 37 manually defined thresholds, (2) logistic regression with L2 regularization, (3) a random forest (500 trees), (4) a standard XGBoost without cost-sensitive modifications, and (5) an isolation forest for unsupervised detection.

6.2. Performance Analysis and Results

Table 1 presents performance comparison on the test set; unless otherwise noted, results are point estimates from a single training run, and stability analysis across random seeds is left for future work. The proposed focal loss ensemble achieved 61.3% precision, 73.2% recall, and 66.7% F1-score, representing +16.0 percentage points precision improvement and +11.1 points recall improvement compared to standard ensemble baseline. AUC-ROC reached 0.918 with false positive rate of 1.35%. Cost-sensitive XGBoost demonstrated marked improvement over standard XGBoost (F1: 60.8% vs 54.1%) through sample weighting. Weighted stacking achieved 62.9% F1-score by combining gradient boosting, neural networks, and graph models.

The legacy rule-based system achieved only 38.2% precision and 51.7% recall (F1=43.9%) with high false positive rate (3.47%), demonstrating substantial performance gap justifying AI-driven approaches. Architecture ablation study validated the proposed multi-stage framework. The complete framework (DFM + GNN + Ensemble) achieved F1-score of 66.7%. Removing GNN features decreased F1 to 60.2% (-6.5 points), demonstrating the value of relational modeling. Using only DFM embeddings without ensemble achieved F1 of 58.4%. Using only traditional hand-crafted features without deep learning achieved F1 of 54.1%. These results confirm that the integration of deep factorization machines, graph neural networks, and ensemble learning provides complementary strengths.

The final deployed model architecture uses: DFM with embedding dimension 64 for categorical features, 3-layer GCN with hidden dimensions [128,64,32] for graph features, 1D-CNN with 64 filters for temporal sequences, cost-sensitive XGBoost (n_estimators = 500, max_depth = 6, scale_pos_weight = 20), focal loss neural network (2 hidden layers [256,128], gamma = 2.0), and meta-ensemble logistic regression (C = 1.0). Total inference time per claim is 12.3ms on CPU (averaged over test set).

Table 3 quantifies individual feature group contributions through ablation studies. Removing visit frequency features decreased F1-score by 4.8 points, billing amount features by 5.2 points, prescription features by 3.7 points, and graph-based relational features by 6.1 points, validating the multi-dimensional approach. Graph neural networks particularly excelled on fraud rings, achieving 71.2% F1-score versus 58.4% for non-graph methods on provider network cases.

SHAP-based importance analysis revealed total 90-day reimbursement as most influential (0.147), followed by provider claim frequency (0.128) and billing z-score (0.116), with top 20 features accounting for 73.4% of total importance.

7. Conclusion and Limitations

This research presented a comprehensive deep learning framework achieving 66.7% F1-score and 73.2% recall, representing 52% and 42% improvements over rule-based systems. The framework integrated multi-dimensional behavioral features, cost-sensitive ensemble learning, and explainable AI mechanisms.

Several limitations warrant consideration. First, the dataset derived from a single regional provider, potentially limiting generalizability to different healthcare contexts. Second, ground truth labels originated from confirmed investigations, creating potential blind spots for undetected sophisticated fraud. Third, evaluation focused on binary classification without distinguishing fraud typologies. Fourth, computational constraints limited graph analysis to radius-2 ego-networks. Fifth, the study evaluated offline performance without prospective deployment measuring real-world operational impact.

Future research directions include transfer learning across healthcare systems, federated learning for privacy-preserving collaborative training, temporal modeling through recurrent architectures, causal inference for distinguishing fraud from policy changes, and integration with external knowledge sources including regulatory databases and medical ontologies.

References

1. 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, p. 103061, 2025. doi: 10.1016/j.artmed.2024.103061
2. N. Kumaraswamy, M. K. Markey, T. Ekin, J. C. Barner, and K. Rascati, "Healthcare fraud data mining methods: a look back and look ahead," *Perspectives in health information management*, vol. 19, no. 1, p. 1i, 2022.
3. R. Pingili, "AI-Powered Claims Intelligence for Identifying Billing Anomalies and Fraud in Medicare and Medicaid," *Journal of Computer Science and Technology Studies*, vol. 7, no. 10, pp. 290-305, 2025.
4. J. T. Hancock, R. A. Bauder, H. Wang, and T. M. Khoshgoftaar, "Explainable machine learning models for Medicare fraud detection," *Journal of Big Data*, vol. 10, no. 1, p. 154, 2023. doi: 10.1186/s40537-023-00821-5
5. 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
6. R. Zhang, D. Cheng, J. Yang, Y. Ouyang, X. Wu, Y. Zheng, and C. Jiang, "Pre-trained online contrastive learning for insurance fraud detection," In *Proceedings of the AAAI Conference on Artificial Intelligence*, March, 2024, pp. 22511-22519. doi: 10.1609/aaai.v38i20.30259
7. J. Lu, B. C. Fung, and W. K. Cheung, "Embedding for anomaly detection on health insurance claims," In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, October, 2020, pp. 459-468.
8. K. Razzaq, and M. Shah, "Next-generation machine learning in healthcare fraud detection: Current trends, challenges, and future research directions," *Information*, vol. 16, no. 9, p. 730, 2025. doi: 10.3390/info16090730
9. H. De Meulemeester, F. De Smet, J. van Dorst, E. Derroitte, and B. De Moor, "Explainable unsupervised anomaly detection for healthcare insurance data," *BMC Medical Informatics and Decision Making*, vol. 25, no. 1, p. 14, 2025. doi: 10.1186/s12911-024-02823-6
10. 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, 2024. doi: 10.1016/j.heliyon.2024.e30045
11. J. M. Johnson, and T. M. Khoshgoftaar, "Data-centric ai for healthcare fraud detection," *SN Computer Science*, vol. 4, no. 4, p. 389, 2023. doi: 10.1007/s42979-023-01809-x
12. L. Settipalli, G. R. Gangadharan, and U. Fiore, "Predictive and adaptive drift analysis on decomposed healthcare claims using ART based topological clustering," *Information Processing & Management*, vol. 59, no. 3, p. 102887, 2022.
13. A. Wahid, M. Msahli, A. Bifet, and G. Memmi, "NFA: A neural factorization autoencoder based online telephony fraud detection," *Digital Communications and Networks*, vol. 10, no. 1, pp. 158-167, 2024. doi: 10.1016/j.dcan.2023.03.002
14. 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, no. 1, p. 62, 2023. doi: 10.1186/s12911-023-02152-0
15. R. Yin, J. Li, Q. Yang, X. Chen, X. Zhang, M. Lin, and A. Subramaniam, "MTLNFM: A Multi-Task Framework Using Neural Factorization Machines to Predict Patient Clinical Outcomes," *Applied Sciences*, vol. 15, no. 15, p. 8733, 2025. doi: 10.1101/2025.05.15.25327659
16. Z. Wang, X. Chen, Y. Wu, L. Jiang, S. Lin, and G. Qiu, "A robust and interpretable ensemble machine learning model for predicting healthcare insurance fraud," *Scientific Reports*, vol. 15, no. 1, p. 218, 2025. doi: 10.1038/s41598-024-82062-x
17. I. Matloob, S. Khan, R. Rukaiya, H. Alfraihi, and J. Ali Khan, "Healthcare fraud detection using adaptive learning and deep learning techniques," *Evolving Systems*, vol. 16, no. 2, p. 72, 2025. doi: 10.1007/s12530-025-09698-6