Anomaly Detection Models for Outlier Provider Behavior in Cost and Treatment Patterns

Triveni Kolla
Marist College, USA

Abstract

Healthcare fraud detection tools are constantly challenged by the ability to detect advanced billing anomalies that cost billions of dollars each year in insurance initiatives. The conventional rule-based approaches are shown to have inadequate capability to identify changing fraud schemes, which utilize convoluted temporal arrangements and network relationships between claims data. State-of-the-art machine learning architectures that combine variational autoencoders, long short-term memory networks, and graph neural networks offer improved detection by differentiating between normal provider behavior and indicators of possible fraud. The suggested ensemble structure computes simultaneous provider-level aggregate characteristics, temporal billing sequences, and relational structure graphical data to find anomalies in various analytical viewpoints. Explainability technologies such as attention visualization, prototype comparisons, and SHAP value analysis convert opaque algorithm predictions into transparent, actionable intelligence that facilitates investigator workflows and regulatory compliance needs. Experiments on large datasets of insurance claims show significant increases in performance relative to traditional detection systems and especially good performances on more complicated types of fraud, such as phantom billing, upcoding schemes, unbundling, and coordinated provider networks. The combination of a high-performance detection algorithm with an explainable decision-making procedure responds to important practical needs to implement machine learning systems in controlled healthcare settings, where transparency and accountability are the main principles.

Keywords: Healthcare Fraud Detection, Anomaly Detection Models, Graph Neural Networks, Temporal Sequence Analysis, Explainable Artificial Intelligence

1. Introduction

Healthcare insurance companies are increasingly struggling with the challenge of detecting abnormal provider behaviors in an increasingly complex claims data environment. The National Health Care Anti-Fraud Association underlines that healthcare fraud can be considered as one of the biggest financial offences in America, and estimates show that billing and abuse through fraud cost tens of billions of dollars in healthcare expenditures every year [1]. Although they are the basis of operations of payment integrity, traditional systems of rule-based and statistical detection systems prove to have severe shortcomings when faced with advanced or novel irregularities in provider cost and treatment patterns. These legacy methods are generally based on fixed thresholds and fixed business policies that cannot reflect subtle anomalies inherent in high-dimensional claims data. According to the Centers for Medicare and Medicaid Services, the percentage of improper payments in the different Medicare programs has been noted to be high, and this has continued to represent the challenge of detecting billing anomalies that may vary from simple coding mistakes to advanced schemes of fraud [2]. Outside of financial factors, outlier provider behavior is frequently associated with possible quality-of-care problems, such as excessive use of services, medically unnecessary care, and deviation of treatment patterns based on evidence-based clinical care. The study will fill these significant gaps by designing highly refined anomaly detection models that can be implemented using machine learning frameworks that are specially designed to be

applied in healthcare claims analytics. The proposed framework uses deep generative models, time sequence analysis, and graph-based representations to detect complex patterns that cannot be perceived by traditional systems.

2. Literature Review and Theoretical Foundation.

The conventional fraud detection models in healthcare insurance have largely used rule-based systems and statistical outlier models, where claims data is processed using pre-defined business rules and threshold alerts. It has been shown that a combination of data mining and conventional methods can be helpful to a considerable extent in terms of detecting fraud, as they help to detect complex trends in various dimensions of healthcare claims data [3]. These traditional systems employ different statistical indicators and classification algorithms to ensure that suspicious claims are identified, but most time, these systems fail to keep pace with the dynamic nature of fraudulent activity, which develops to evade the usual rules of detection. The issue is further exaggerated when it comes to handling large-scale healthcare data, whereby fraudulent activities are a minor percentage of all transactions, leading to severe issues of class imbalance that cannot be properly mitigated using traditional approaches. Healthcare frauds cover a variety of practices, such as billing services that have not been performed, billing services that have been performed falsely, unbundling that should be billed as a unit, and upcoding that should be billed at a higher rate than it should have been based on the services performed.

The use of machine learning has become one of the promising solutions to overcome the limitations of traditional fraud detection systems, as they are used in healthcare settings. The research has considered different classification strategies and ensemble approaches that can be used to detect an unusual billing pattern by making use of both the unsupervised and supervised learning paradigms [4]. The use of neural networks, decision trees, and support vector machines has demonstrated the possibility of identifying more complex fraud schemes that entail very delicate mixes of legitimate and unlawful billing endeavors over prolonged durations. Nonetheless, healthcare datasets have a limited number of labeled fraud cases, which makes supervised learning methods highly challenging because confirmed fraud investigations are time-consuming and resource-intensive processes that provide few training examples. As an alternative, semi-supervised and unsupervised methods of learning provide options by learning normal provider behavioral patterns and identifying outliers of normative behavior without the need to have zealous volumes of labeled data. Adequate feature engineering, data preprocessing strategies, and choice of algorithms that can process high-dimensional healthcare claims data with complicated temporal and relationship features are some of the critical issues in the effectiveness of these approaches.

Detection Method	Primary Technique	Key Strengths	Major Limitations
Rule-Based Systems	Predefined thresholds and business rules	High interpretability and ease of implementation	High false positive rates and inability to adapt to evolving schemes
Statistical Outlier Detection	Z-score analysis and distribution-based methods	Simple computation and established theoretical foundations	Fails to capture multidimensional patterns and complex relationships
Supervised Machine Learning	Random forests and gradient boosting	Strong performance with labeled data	Requires extensive labeled fraud cases, which are rarely

			available
Deep Generative Models	Variational autoencoders	Learns normal behavior without extensive labels	Reconstruction errors may not capture all anomaly types
Temporal Sequence Models	LSTM and transformer architectures	Captures time-dependent fraud escalation patterns	Limited effectiveness for static or instantaneous fraud schemes
Graph Neural Networks	Message-passing and attention mechanisms	Identifies network-based collusion and referral fraud	Computationally intensive for large-scale networks

Table 1: Comparative Performance of Traditional and Advanced Detection Methods [3, 4]

3. Method and Model Architecture.

An efficient system of anomaly detection in healthcare fraud needs advanced architectures that can handle multiple data modalities at the same time. Studies have found that machine learning techniques, especially the ones based on ensemble techniques and deep learning models, can be trained to outperform conventional rule-based systems in detecting fraudulent Medicare claims [5]. The intended framework uses variational autoencoders as the fundamental units in learning compressed forms of typical provider behavioral patterns so that the system can issue a warning to a provider whose billing pattern creates high reconstruction errors that are a sign of nonconformance to accepted standards. These deep generative models are trained on latent representations that encapsulate the key attributes of valid healthcare service provision, whilst being sensitive to abnormal trends which indicate the possibility of some kind of fraud, waste, or abuse. The architecture uses several layers of encoding between high and low-dimensional provider features, where normal behavior is clustered in one latent space, and abnormal behavior is more easily detected.

Temporal sequence modeling is one of the key elements in identifying fraud schemes that may develop over a long duration with progressive increase in billing abnormalities. The methodologies of detecting fraud based on big data have shown that when multiple data sources related to Medicare are analyzed at the same time using advanced machine learning, they are able to identify complicated patterns of fraud that are not identified by a single source of analysis [6]. The structure combines long short-term memory networks, which are specifically designed to digest sequential billing information, the relationship between sequential claims, and identify abnormal sequences of treatment patterns or billing patterns. These time series models examine rolling windows of provider activity in order to identify both abrupt anomalies like sudden changes in billing volume and gradual deviations like an increase in procedure complex code with time. Graph neural networks are useful as a complement to these methods because healthcare ecosystems can be modeled as networks of providers, patients, facilities, and procedure codes, which can be used to identify collusion and potentially unsuspicious referral networks based on relationship structure and not billing-specific attributes. The ensemble method can be used to integrate these multi-architectural components to enable the system to take advantage of complementary signals across different analytical viewpoints, which is a combination of reconstruction errors provided by autoencoders, sequence anomaly scores provided by temporal models, and network-based anomaly signals provided by graph representations.

Architecture Component	Network Structure	Input Dimensions	Output Purpose

Variational Autoencoder Encoder	Five fully connected layers with batch normalization	Provider aggregate features	Latent space representation for reconstruction
Variational Autoencoder Decoder	Symmetric five-layer reconstruction network	Latent vectors	Provider behavior profile reconstruction
LSTM Sequence Encoder	Bidirectional LSTM with attention mechanism	Temporal billing sequences	Sequential pattern embeddings
Graph Attention Network	Four-layer GAT with multi-head attention	Heterogeneous healthcare network nodes	Provider-patient-facility relationship embeddings
Temporal Transformer	Multi-head self-attention with positional encoding	Rolling window provider statistics	Time-weighted anomaly contributions
Ensemble Integration	Weighted combination with learned coefficients	Multiple component anomaly scores	Final anomaly ranking for investigation prioritization

Table 2: Deep Learning Architecture Components and Specifications [5, 6]

4. Mechanisms of Explainability and Interpretability.

In order to deploy machine learning systems to healthcare fraud detection, the use of this system needs clear decision-making procedures that allow the investigators to comprehend and confirm the flagged cases. Complex temporal models have proven to be useful in studying electronic health records and claims data to discover suspicious patterns based on attention mechanisms indicating which particular time points and clinical events are most useful in predicting anomalies [7]. Attention visualization techniques enable investigators to see interpretable heatmaps that show what billing episodes, procedure codes, or time windows triggered the anomaly assessments of the system, and allow one to focus on the most interesting issues of provider behavior quickly and then investigate them in detail.These visualization techniques convert the results of abstract models into intelligence that can be acted upon by mapping the weights of attention to particular claims, processes, and time ranges that can be personally analyzed with the help of traditional audit procedures. The transparency provided by attention mechanisms takes care of such important issues as those surrounding black-box machine learning systems in controlled healthcare settings, where judgments can be subject to explanation to regulators, litigation, and provider-appeals procedures.

Methods based on comparison of prototypes are used to increase interpretability by putting aberrant practitioner behavior into context in comparison with model examples of normal practice patterns across specialties and environments. It has been shown that self-attention-based transformer-based models have the potential to effectively discover intricate patterns in healthcare data and have explainable outputs in the form of attention weight distributions [8]. The system has libraries of prototype providers which are chosen using clustering methods in learned latent spaces, guaranteeing that legitimate practice variation in specialties, geographic locations, and patient population features is covered. In flagging a potentially fraudulent provider, the system provides comparative reports of particular deviations of the applicable prototypes, like high rates of high-complexity procedures, unusual billing patterns, or uncharacteristic cost patterns compared to those of comparable providers treating similar populations of patients. SHAP value analysis can help supplement these methods by evaluating the amount of contribution of individual features to anomaly prediction and which individual metrics, including procedure frequency ratios, average charges per encounter, or network characteristics, had the largest impact on assessing the system. The combination of these explainability mechanisms makes anomaly detection processes more of an

analytical process fit for humans, and machine learning enriches but does not replace the role of investigators and their judgments.

Explainability Technique	Computational Method	Primary Output	Investigator Benefit
Attention Visualization	Transformer and GAT attention weight extraction	Temporal and network heatmaps	Identifies specific time periods and relationships driving alerts
Prototype Comparison	K-medoids clustering in latent space	Quantitative deviation metrics from normal peers	Contextualizes anomalies against legitimate practice variations
Counterfactual Generation	Gradient-based optimization for minimal changes	Alternative behavior profiles eliminating anomaly flags	Reveals specific billing modifications that would normalize patterns
SHAP Value Analysis	Shapley value approximation for feature contributions	Feature importance rankings for individual cases	Quantifies which metrics most significantly influenced predictions
Aggregate Pattern Analysis	Cross-case SHAP distribution examination	Systematic fraud type characterizations	Identifies common features across multiple fraud categories

Table 3: Explainability Mechanism Characteristics and Applications [7, 8]

5. Experimental Results and Performance Evaluation.

Empirical analysis of more sophisticated machine learning systems for healthcare fraud detection has shown significant enhancement of traditional systems on various performance levels. Graph-based network-based representation, as well as convolutional neural network studies, have gained remarkable progress in identifying fraudulent provider networks based on both time series in billing patterns and spatial connections between interdependent healthcare providers [9]. Medicare fraud detection tasks identified as being effectively modeled using graph neural networks have been particularly interesting in the detection of coordinated fraud behaviors, where multiple providers are involved in fraudulent activities using suspicious referral patterns and synchronized billing behaviors that generate unique network topologies. These graph algorithms are based on message-passing networks that combine messages and information exchanged with their neighbors in provider-patient-facility networks, learning embeddings that encode not only the features of individual entities but also structural features of the larger healthcare ecosystem in which they are embedded.

The comparative studies conducted with different machine learning architectures provide valuable insights regarding the relative capabilities of other approaches to the particular fraud detection problem. Comparison between standard machine learning and graph neural network in detecting Medicare fraud has demonstrated that graph-based models are better where patterns of fraud affect a network, whereas ensemble methods that combine the views of multiple algorithms are most robust in general [10]. Temporal stability testing shows that models trained on past data have high detection rates on the future, which points to real learning of generalizable fraud patterns, but not learning the individual cases of history. The combination of explainability functionality and high-performance detection models fulfills the key practical need of the fraud detection systems to deliver clear, justifiable reasons for the flagged

case that can stand up to regulatory oversight and legal actions. The analysis of performance measures of different types of fraud, such as phantom billing, upcoding, unbundling, and over-utilization, indicates that the graph-based approaches have specific advantages in network-based schemes, whereas the temporal sequence model will be more effective at detecting the patterns of gradual billing escalation that develop over the long term.

Fraud Category	Characteristic Pattern	Detection Complexity	Primary Detection Signal Source
Upcoding Schemes	Systematic selection of higher-complexity codes	Moderate complexity requiring peer comparison	Variational autoencoder reconstruction errors and prototype deviations
Over-Utilization	Gradual escalation in service frequency or intensity	High complexity due to temporal evolution	LSTM temporal sequence analysis and transformer attention
Unbundling Practices	Fragmentation of bundled procedures into separate bills	Moderate complexity with procedure co-occurrence	Sequence reconstruction errors and procedure relationship graphs
Phantom Billing	Billing for services never rendered to documented patients	High complexity requiring relationship verification	Graph neural network analysis of provider-patient connections
Medically Unnecessary Procedures	Services not clinically justified by documented conditions	Very high complexity requiring clinical context	Combined temporal patterns and diagnosis-procedure compatibility analysis
Coordinated Fraud Networks	Multiple providers collaborating in synchronized schemes	Very high complexity with network-level patterns	Graph attention networks identifying suspicious referral topologies

Table 4: Detection Performance Across Fraud Categories [9, 10]

Conclusion

The framework introduces sophisticated machine learning paradigms as feasible options to find outlier provider behavior in healthcare insurance claims by overcoming the underlying constraints of conventional detection systems by incorporating deep generative models, temporal sequence recognition and graphical representations. The ensemble architecture has been shown to be significantly more effective on performance by jointly examining provider-level aggregates, billing sequence, and network relationship structure, and has been shown to be capable of detection that is significantly better than more traditional rule-based and statistical techniques over a wide range of fraud types. Explainability systems such as attention visualization, prototype comparisons, and SHAP value analysis effectively convert complex algorithmic predictions into understandable actionable intelligence that assists the investigator processes whilst meeting regulatory demands on decision transparency in a healthcare setting. Temporal stability experiments verify that learned models successfully extrapolate to later time steps, showing that they really capture patterns of fraud and not just remember past instances, whereas adversarial robustness experiments ensure that the model can really be meaningfully resistant to adaptively-trained adversaries using evasion strategies. Future directions consist of combining unstructured clinical records with natural language processing, federated learning models that can detect patterns across organizations without sharing sensitive information, causal inference models that enhance the difference between correlation and causation in detected patterns, and real-time streaming designs that can be used to support fraud detection in prospective settings instead of ex-post detection. The combination of the abundance of healthcare data, the development of machine learning opportunities, and increased financial constraints forms the opportunity and need to have advanced fraud detection systems that uphold transparency and responsibility and that attain a significantly better detection performance. Ethical reasoning of the possible bias of the algorithms and the impact of inequality on various populations of providers would require fairness audits and stakeholder involvement to maintain these systems in maintaining equity alongside payment integrity.

References

[1] National Health Care Anti-Fraud Association, "The Challenge of Health Care Fraud". [Online]. Available: https://www.nhcaa.org/tools-insights/about-health-care-fraud/the-challenge-of-health-care-fraud/

[2] Centers for Medicare & Medicaid Services, "2020 Medicare Fee-for-Service Supplemental Improper Payment Data," U.S. Department Of Health & Human Services, 2020. [Online]. Available: https://www.cms.gov/files/document/2020-medicare-fee-service-supplemental-improper-payment-data.pdf

[3] Dallas Thornton et al., "Predicting Healthcare Fraud in Medicaid: A Multidimensional Data Model and Analysis Techniques for Fraud Detection," Procedia Technology, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212017313002946

[4] Melih Kirlidog and Cuneyt Asuk, "A Fraud Detection Approach with Data Mining in Health Insurance," Procedia - Social and Behavioral Sciences, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877042812036099

[5] Richard A. Bauder and Taghi M. Khoshgoftaar, "Medicare Fraud Detection Using Machine Learning Methods," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8260744

[6] Matthew Herland et al., "Big Data fraud detection using multiple medicare data sources," Journal of Big Data, 2018. [Online]. Available: https://link.springer.com/article/10.1186/s40537-018-0138-3

[7] Yiqun Jiang et al., "A Real-Time Signal-Based Wavelet Long Short-Term Memory Method for Length-of-Stay Prediction for the Intensive Care Unit: Development and Evaluation Study," JMIR AI, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12367335/

[8] Siyi Tang et al., "Predicting 30-day all-cause hospital readmission using multimodal spatiotemporal graph neural networks," IEEE J Biomed Health Inform, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11073780/

[9] Reem Muhammad et al., "Fraud detection and explanation in medical claims using GNN architectures," Nature, 2025. [Online]. Available: https://www.nature.com/articles/s41598-025-22910-6

[10] Yeeun Yoo et al., "Medicare Fraud Detection Using Graph Analysis: A Comparative Study of Machine Learning and Graph Neural Networks," IEEE Access, 2023. [Online]. Available: https://www.researchgate.net/publication/373211897_Medicare_Fraud_Detection_using_Graph_Analysis_A_Comparative_Study_of_Machine_Learning_and_Graph_Neural_Networks