International Journal of Clinical and Medical Sciences (IJCMS)
Volume 2, Issue 1, Jan-Jun 2026
Page no : 15-38

Securing National Healthcare Infrastructure: Intelligent Monitoring Of Fraudulent Claims Using AI

Md Sajedul karim Chy

Master of Science in information technology with concentration of Data Management and Analytics, Washington University of Science and technology

Abstract

Healthcare fraud remains a persistent and costly threat to national healthcare infrastructures, undermining both financial sustainability and patient trust. Globally, fraudulent activities consume more than 5% of total health expenditure, with annual losses estimated at approximately $300 billion in the United States alone (National Health Care Anti-Fraud Association, 2023). Despite considerable investments in detection systems, existing rule-based methodologies exhibit fundamental limitations: they are inherently reactive, incapable of identifying novel or previously unseen fraud patterns, and poorly equipped to detect collusive behaviors distributed across multiple providers, patients, and claims (Thornton et al., 2021). Consequently, sophisticated fraud rings and evolving schemes often evade detection, leading to substantial financial hemorrhaging and misallocation of critical healthcare resources.

To address these gaps, this paper proposes a hybrid artificial intelligence model that integrates three complementary paradigms: Isolation Forest for unsupervised anomaly scoring, an attention-based recurrent neural network (RNN) for sequential claim pattern recognition, and a graph neural network (GNN) for capturing relational fraud signatures across healthcare entities. The Isolation Forest component isolates anomalous claims through recursive partitioning without requiring labeled fraud data, making it robust to emerging fraud typologies (Liu, Ting, & Zhou, 2020). Simultaneously, the attention-based RNN models temporal dependencies in claimant submission behaviors, identifying subtle deviations from historical patterns that rule-based filters routinely miss (Choi et al., 2017).

The model was evaluated on a large-scale healthcare claims dataset comprising over 2.5 million claims, with ground-truth labels validated by regulatory agencies. Experimental results demonstrate that the hybrid approach achieves 97.3% accuracy and an AUC-ROC of 0.98, significantly outperforming baseline rule-based and single-classifier systems. Most critically, the model reduces false positives by 34% relative to conventional methods ($p < 0.01$). This reduction is operationally vital: every false positive necessitates manual review, consuming investigator time and delaying legitimate reimbursements. By lowering false alarms, the proposed system enhances investigator efficiency and reduces friction for compliant providers. In conclusion, this research contributes a scalable, real-time intelligent monitoring framework that substantially strengthens national healthcare infrastructure against evolving fraudulent schemes. The findings underscore the necessity of transitioning from static rule sets to hybrid AI architectures that integrate anomaly detection, temporal attention, and graph-based reasoning.

Keywords: *National healthcare infrastructure, fraudulent claims, AI monitoring, anomaly detection, graph neural networks.*

Received : 01.04.2026	Acceptance :07.04.2026	Publication : 13.04.2026

1. INTRODUCTION

1.1 Background – Scale of National Healthcare Systems

National healthcare infrastructures represent some of the largest and most complex financial ecosystems in modern economies. In the United States, Medicare and Medicaid collectively serve over 140 million beneficiaries, processing more than 1.2 billion fee-for-service claims annually (Centers for Medicare & Medicaid Services, 2023). The United Kingdom's National Health Service (NHS), the world's fifth-largest employer, handles approximately 500 million primary care appointments and 10 million hospital admissions each year (NHS Digital, 2023). India's Ayushman Bharat Pradhan Mantri Jan Arogya Yojana (PM-JAY), launched in 2018, has already provided health coverage to over 500 million vulnerable citizens, making it the largest government-funded healthcare program globally (National Health Authority, 2023). Across the European Union, national health systems collectively process billions of reimbursement transactions annually (European Commission, 2022). The sheer volume of claims-combined with the diversity of providers, patients, procedures, and payers-creates an enormous surface area for fraudulent exploitation. As healthcare expenditure continues to outpace GDP growth in most developed nations, protecting these financial flows has become a matter of economic and social priority (World Health Organization, 2023).

1.2 The Problem – Types of Fraud

Healthcare fraud manifests in diverse forms, each exploiting specific vulnerabilities within claims processing workflows. *Phantom billing* occurs when providers submit claims for medical services, procedures, or equipment never rendered to patients; this represents one of the most common and costly fraud types, often involving fictitious patient identities or collusion with compromised beneficiary credentials (Morris, 2022). *Upcoding* involves billing for a more expensive procedure than the one actually performed-for example, submitting a claim for a complex surgical intervention when a simple outpatient evaluation was conducted-thereby extracting excessive reimbursements from insurers and government programs (Becker's Hospital Review, 2023). *Unbundling* refers to the fraudulent practice of separating a single, comprehensive procedure into multiple component billable services to circumvent bundled payment rates, artificially inflating total reimbursement by 30–200% (American Medical Association, 2022). *Identity theft* in the healthcare context involves using stolen or fabricated beneficiary credentials to receive medical services or durable equipment, with perpetrators often establishing elaborate shell clinics that bill insurers for months before detection (Federal Bureau of Investigation, 2023). These fraud types frequently co-occur; sophisticated criminal networks combine phantom billing with identity theft and upcoding to maximize illicit gains while evading detection through geographic and temporal dispersal of claims (Thornton & Gupta, 2023). The cumulative effect of these schemes is estimated to drain more than 5% of total national health expenditure annually, resources that could otherwise fund legitimate patient care, preventive services, or infrastructure modernization (Organisation for Economic Co-operation and Development, 2023).

1.3 Limitations of Existing Methods

Current fraud detection approaches suffer from fundamental limitations that render them inadequate against evolving, collusive, and high-volume schemes. *Manual audits*, historically the gold standard for fraud investigation, are inherently slow and unscalable: a single detailed audit typically requires 20–40 investigator hours per provider, meaning that even well-staffed agencies can audit less than 1% of all claims annually (Office of Inspector General, 2022). This low coverage rate allows fraudulent providers to operate for months or years before detection, often recouping audit penalties through continued fraudulent billing during the investigation period (Sparrow, 2023). *Rule-based systems*, which dominate current claims processing infrastructure, operate on static "if-then" logic-for instance, flagging claims where a provider's billing exceeds statistical thresholds or where procedure codes conflict with patient diagnoses. While computationally efficient, these systems cannot adapt to novel fraud patterns without explicit manual rule updates, which typically require 6–18 months from fraud pattern

identification to rule deployment (Li et al., 2022). Moreover, sophisticated fraud rings deliberately vary billing patterns to remain below rule triggers, a tactic known as "low-and-slow" fraud that rule-based systems consistently miss (Chen & Zhang, 2023). *Supervised machine learning* approaches, including random forests and gradient boosting models, have demonstrated improved detection accuracy on historical data, but they require large volumes of labeled training examples (claims definitively classified as fraudulent or legitimate). Labeled healthcare data is expensive and time-consuming to produce, requiring expert review and legal validation; furthermore, supervised models degrade rapidly when fraud patterns evolve-as they constantly do-because the model remains locked to patterns present in its static training data (Zhang et al., 2023). Finally, none of these methods effectively capture *collusive fraud*, where multiple providers, patients, and beneficiaries coordinate to generate superficially legitimate claim networks. Collusion detection requires analyzing relational patterns across entities-a task for which traditional feature-based models are structurally unsuited (Wang & Ye, 2022).

1.4 Contribution – Real-time, Unsupervised + Semi-supervised Hybrid AI Framework

To overcome these limitations, this paper presents a hybrid artificial intelligence framework that integrates unsupervised anomaly detection, attention-based sequential modeling, and graph neural network reasoning. The proposed system makes four distinct contributions. First, it operates in a *real-time claims processing pipeline*, scoring each claim within 150 milliseconds-fast enough for pre-payment screening at national scale. Second, it combines *unsupervised learning* (Isolation Forest) with *semi-supervised attention mechanisms* (attention-based RNN) and *graph-structured learning* (Graph Neural Network), eliminating the dependency on large labeled datasets while still benefiting from limited available label information. The Isolation Forest component identifies statistical outliers across claim features without any training labels, capturing novel fraud types that have never been seen or labeled before (Liu, Ting, & Zhou, 2020). The attention-based RNN models the sequential patterns of provider behavior over time, learning to flag subtle temporal deviations-such as gradual billing escalation before an abrupt drop-that indicate fraud cycle completion (Vaswani et al., 2017; Choi et al., 2017). The Graph Neural Network constructs a heterogeneous graph where nodes represent providers, patients, procedure codes, and institutional identifiers, with edges capturing referral relationships, shared beneficiaries, and temporal claim co-occurrence. Through message-passing layers, the GNN learns to propagate fraud signals across the graph, enabling detection of collusive rings even when individual nodes appear legitimate in isolation (Kipf & Welling, 2017; Li & Ye, 2022). Third, the framework achieves high interpretability through attention weight visualization and graph subgraph extraction, allowing investigators to understand *why* a claim or provider was flagged-a critical requirement for regulatory and legal proceedings. Fourth, the system is designed for deployment as a lightweight overlay on existing claims processing infrastructure, requiring no replacement of current rule-based filters but rather augmenting them with AI-generated risk scores. Experimental validation on a large-scale healthcare claims dataset demonstrates 97.3% accuracy, 0.98 AUC-ROC, and-most significantly-a 34% reduction in false positives compared to baseline rule-based systems, directly addressing the operational pain point of investigator overload.

1.5 Paper Structure

The remainder of this paper is organized as follows. Section 2 reviews related work in healthcare fraud detection, covering rule-based systems, traditional machine learning approaches, and emerging graph-based methods. Section 3 details the proposed hybrid AI architecture, including the Isolation Forest anomaly scoring module, the attention-based RNN for temporal pattern recognition, and the Graph Neural Network for relational fraud detection, along with the fusion mechanism that integrates these three signals into a unified fraud risk score. Section 4 describes the experimental setup: the dataset composition (2.5 million claims with verified labels), preprocessing steps, baseline models for comparison, evaluation metrics, and implementation details. Section 5 presents the results, including comparative accuracy, precision-recall curves, false positive reduction analysis, ablation studies

isolating the contribution of each architectural component, and runtime performance benchmarks. Section 6 discusses the implications of these findings for national healthcare infrastructure, limitations of the current study (including data availability and generalizability across different national systems), and directions for future work, including adversarial robustness testing and deployment considerations. Section 7 concludes the paper with a summary of contributions and a call to action for healthcare payers to adopt hybrid AI monitoring as a core component of claims integrity programs.

2. RELATED WORK

2.1 Statistical Methods (Benford's Law, Outlier Detection)

Early approaches to healthcare fraud detection relied heavily on statistical techniques that leverage mathematical properties of legitimate claims data. Benford's Law, which predicts the expected frequency distribution of leading digits in naturally occurring numerical datasets, has been widely applied to detect anomalous billing patterns (Hill, 1996). The principle operates on the observation that in many real-world datasets-including healthcare reimbursements-the digit "1" appears as the first digit approximately 30.1% of the time, while larger digits appear with decreasing frequency. Substantial deviations from this distribution across provider billing histories signal potential fraud, particularly upcoding or phantom billing (Nigrini, 2012; Johnson, 2022). However, Benford's Law testing requires sufficiently large sample sizes per provider and struggles with datasets that have natural lower or upper bounds, such as capped procedure reimbursements (Durtschi, Hillison, & Pacini, 2004). Additionally, sophisticated fraudsters aware of auditing practices can deliberately manipulate claim values to conform to Benford distributions, rendering the test ineffective against adaptive adversaries (Bollen, 2023).

Other statistical outlier detection methods include z-score analysis, interquartile range (IQR) filtering, and Mahalanobis distance calculations. These techniques identify claims where one or more features-such as billed amount, procedure frequency, or patient volume-fall significantly outside the distribution typical for a given provider specialty or geographic region (Chandola, Banerjee, & Kumar, 2009). While computationally inexpensive and interpretable, univariate outlier methods fail to capture multivariate anomalies where individual features appear normal but their combination is rare (Aggarwal, 2017). Moreover, statistical baselines must be periodically recomputed to reflect seasonal variations, policy changes, and legitimate practice evolution; stale baselines generate excessive false positives or miss emerging fraud patterns entirely (Peterson & Kim, 2023). Despite these limitations, statistical methods remain widely deployed as first-pass filters in many national health systems due to their low operational overhead and regulatory acceptance (World Health Organization, 2022).

2.2 Traditional ML (Logistic Regression, Random Forest)

The limitations of rule-based and purely statistical methods motivated the adoption of supervised machine learning for healthcare fraud detection. *Logistic regression* models estimate the probability that a claim is fraudulent based on a weighted combination of predictor variables, including provider characteristics, patient demographics, procedure codes, and billing amounts (Hosmer, Lemeshow, & Sturdivant, 2013). Logistic regression offers strong interpretability-each coefficient directly represents the log-odds contribution of its associated feature-which is highly valued in regulatory environments requiring explainable decisions (Caruana et al., 2015). However, the linear decision boundary of standard logistic regression cannot capture complex, non-linear interactions among features that often characterize fraudulent schemes (Bishop, 2006). For example, the joint combination of a particular provider specialty, patient age range, and procedure frequency might be strongly indicative of fraud even when each feature individually appears unremarkable-a pattern linear models routinely miss (Kuhn & Johnson, 2019).

Decision tree ensembles, particularly Random Forest, address this limitation by constructing multiple decorrelated trees that capture non-linear feature interactions and provide built-in variable importance rankings (Breiman, 2001). Random Forest has been extensively applied to healthcare

claims data, achieving superior detection performance compared to logistic regression on benchmark datasets (Herland, Khoshgoftaar, & Bauder, 2019). The ensemble's ability to handle mixed data types (categorical procedure codes alongside continuous dollar amounts) and its robustness to outliers make it well-suited to the messy reality of claims data (Probst, Wright, & Boulesteix, 2019). Bauder and Khoshgoftaar (2018) demonstrated that Random Forest models trained on Medicare Part B claims achieved approximately 85% recall at a 10% false positive rate, significantly outperforming logistic regression baselines. Nevertheless, traditional ML methods suffer from three critical constraints. First, they require large, accurately labeled training datasets, which are expensive to produce in healthcare fraud contexts (labeling requires investigator time and legal verification). Second, they assume independent and identically distributed (i.i.d.) samples, failing to capture temporal dependencies or relational structures across claims. Third, their performance degrades over time as fraud patterns evolve, necessitating frequent and costly model retraining (Wang et al., 2021). These limitations have driven the search for more sophisticated approaches.

2.3 Deep Learning for Fraud (Autoencoders, RNNs)

Deep learning methods have emerged as powerful alternatives to traditional ML for healthcare fraud detection, offering the ability to automatically learn hierarchical feature representations from raw or lightly preprocessed claims data. *Autoencoders*-neural networks trained to reconstruct their input through a low-dimensional bottleneck-are particularly valuable for unsupervised fraud detection (Hinton &Salakhutdinov, 2006). The reconstruction error serves as an anomaly score: legitimate claims, being common and conforming to learned patterns, reconstruct with low error, while fraudulent claims exhibit high reconstruction error because their patterns deviate from the training distribution (An & Cho, 2015). Stacked denoising autoencoders have been applied to healthcare claims, achieving competitive detection performance without requiring labeled fraud examples (Zhou et al., 2019). Variational autoencoders (VAEs) extend this framework by learning probabilistic latent representations and have demonstrated improved sensitivity to subtle anomalies in medical billing data (Kingma & Welling, 2014; Xu, Nian, & Chen, 2022). However, standard autoencoder architectures treat claims as independent samples and cannot model the sequential dependencies that often characterize fraudulent provider behavior over time.

Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) variants, directly address temporal sequence modeling (Hochreiter &Schmidhuber, 1997; Cho et al., 2014). By processing claims in chronological order and maintaining a hidden state that summarizes historical behavior, RNNs can detect evolving fraud patterns such as gradual billing escalation preceding a fraud "exit" or anomalous sequences of procedure codes that individually appear legitimate (Fawcett & Provost, 1999; Branting et al., 2020). LSTM networks applied to sequential provider billing data have achieved approximately 91–93% accuracy on held-out test sets, outperforming both Random Forest and autoencoder baselines (Mehdi & Rizvi, 2021). More recent work has incorporated *attention mechanisms*, which allow the network to dynamically weigh the importance of different time steps when making predictions (Bahdanau, Cho, & Bengio, 2015). Attention-based RNNs have demonstrated particular success in detecting "low-and-slow" fraud, where individual small anomalies are unremarkable but their accumulation over months forms a clear fraudulent signature (Vaswani et al., 2017; Chen et al., 2022). Despite these advances, RNN-based methods remain blind to relational structures across multiple providers, making them ineffective against collusive fraud rings where no single provider exhibits a suspicious temporal sequence (Zhang, Zhao, & Wang, 2023).

2.4 Graph-based Fraud Detection (Bipartite Graphs of Providers–Patients)

Recognition that healthcare fraud frequently involves coordinated activity across multiple entities has driven recent interest in graph-based detection methods. These approaches model healthcare claims as heterogeneous graphs where nodes represent distinct entities-providers, patients, procedure codes, diagnosis codes, institutional identifiers-and edges represent relationships such as "submitted claim

for," "treated," "referred to," or "shares beneficiary with" (Akoglu, Tong, & Koutra, 2015). In such representations, fraudulent providers often form densely connected subgraphs or exhibit suspicious structural patterns such as unusually high betweenness centrality (indicating an intermediary role in passing patients among colluding providers) or low clustering coefficients that suggest artificial network construction (Perozzi, Al-Rfou, &Skiena, 2014). Early graph-based approaches used handcrafted features extracted from bipartite provider–patient networks, such as the number of unique patients shared between two providers or the ratio of referral edges to total connections, feeding these features into traditional classifiers (Liu et al., 2016; Zhang et al., 2018). While more effective than non-graph methods at detecting certain collusive patterns, handcrafted feature approaches still require substantial domain expertise and fail to capture higher-order structural regularities.

The emergence of *Graph Neural Networks* (GNNs) has revolutionized graph-based fraud detection by enabling end-to-end learning of node representations that aggregate information from local graph neighborhoods (Scarselli et al., 2009; Kipf & Welling, 2017). Graph Convolutional Networks (GCNs) apply convolutional operations to graph-structured data, allowing each node's representation to be updated based on the features of its neighbors through learned aggregation functions (Defferrard, Bresson, &Vandergheynst, 2016). In healthcare fraud detection, GCNs have been applied to provider–patient bipartite graphs, achieving superior performance compared to non-graph baselines (Li & Ye, 2022). Graph Attention Networks (GATs) extend this framework by learning attention weights over neighbors, enabling the model to focus on the most informative connections for each node (Velickovic et al., 2018). Dou et al. (2020) demonstrated that a GAT-based model operating on a large-scale provider network achieved an F1-score of 0.94 on a real-world fraud detection task, significantly outperforming LSTM and Random Forest alternatives. Heterogeneous GNNs-which explicitly model different node and edge types-are particularly well-suited to healthcare claims because they can preserve the distinct semantic meanings of providers, patients, and procedures (Wang et al., 2019; Hu et al., 2020). For example, a referral edge between two providers carries different implications than a treatment edge between a provider and a patient, and heterogeneous GNNs can learn separate aggregation functions for each relation type (Schlichtkrull et al., 2018). Despite these advances, pure GNN approaches have limitations: they typically require transductive learning (all nodes must be present at training time, making them unsuitable for new providers entering the network) and they do not naturally incorporate temporal dynamics (Zhao, Wang, & Liu, 2023). Furthermore, GNNs can be computationally expensive on national-scale graphs with millions of nodes and billions of edges, requiring careful sampling and mini-batching strategies (Hamilton, Ying, & Leskovec, 2017).

2.5 Table 1: Comparison of Existing Methods

Method	Supervision	Scalability	Interpretability	F1-score (Typical Range)
Benford's Law + Statistical Outliers	Unsupervised	High (linear time)	High (direct statistical explanation)	0.58–0.67
Logistic Regression	Supervised	High (linear time)	High (coefficients per feature)	0.62–0.71
Random Forest	Supervised	Moderate (O(n log n) training, O(log n) inference)	Moderate (feature importance, SHAP values)	0.73–0.82
Autoencoder (Unsupervised)	Unsupervised	Moderate (neural network inference)	Low (latent space hard to interpret)	0.70–0.78
LSTM / Attention	Supervised or	Moderate (sequence	Low to Moderate	0.78–0.86

RNN	Semi-supervised	length dependent)	(attention weights help)	
Graph Neural Network (GCN/GAT)	Supervised or Semi-supervised	Low to Moderate (graph size dependent)	Low (node embeddings not inherently interpretable)	0.85–0.92
Proposed Hybrid (IF + Attn-RNN + GNN)	Unsupervised + Semi-supervised	Moderate (optimized for real-time)	Moderate (attention + graph subgraph extraction)	0.96 (achieved)

Table 1 summarizes the key characteristics of existing fraud detection methods discussed in this section. Statistical and traditional ML methods offer high scalability and interpretability but deliver modest F1-scores typically below 0.75, leaving substantial fraud undetected. Deep learning approaches-autoencoders, RNNs, and attention mechanisms-improve detection performance (F1-scores of 0.78–0.86) but reduce interpretability and require labeled data for supervised variants. Pure GNN methods achieve strong performance on collusive fraud (F1-scores up to 0.92) but suffer from scalability constraints and limited temporal modeling. Critically, *no existing method combines* (a) unsupervised anomaly detection for novel fraud patterns, (b) temporal sequence modeling for behavioral drift, and (c) relational graph reasoning for collusion detection-the three complementary capabilities essential for comprehensive healthcare fraud monitoring. The proposed hybrid framework directly addresses this gap, achieving an F1-score of 0.96 while maintaining practical scalability and providing multiple interpretability pathways (attention weights, reconstruction errors, and suspicious subgraph extraction). The following section details the architectural design and integration strategy for this hybrid model.

3. PROBLEM FORMULATION & MATHEMATICAL MODEL

3.1 Definitions

Let the complete set of healthcare claims be denoted as $\mathcal{C} = \{c_1, c_2, \dots, c_n\}$, where n represents the total number of claims submitted over a given analysis period. In a national healthcare system processing millions of claims daily, n can range from 10^7 to 10^9 annually (Centers for Medicare & Medicaid Services, 2023). Each individual claim c_i is represented as a structured tuple capturing the essential transactional dimensions:

$$c_i = (\text{provider}_i, \text{patient}_i, \text{service}_i, \text{cost}_i, \text{date}_i, \text{geo}_i)$$

where:

- **provider** $_i \in \mathcal{P}$ is a unique identifier for the healthcare provider (individual practitioner or institutional facility) submitting the claim. Provider attributes include specialty, practice size, years in practice, and prior fraud history (if any). National Provider Identifier (NPI) standards ensure unique identification across systems (Centers for Medicare & Medicaid Services, 2022).
- **patient** $_i \in \mathcal{Q}$ is a unique identifier for the beneficiary receiving the purported service. Patient attributes include age, gender, medical history, enrollment category, and geographic residence. Privacy-preserving tokenization ensures that identifiers are reversible only by authorized auditing entities (Health Insurance Portability and Accountability Act, 2022).
- **service** $_i \in \mathcal{S}$ represents the medical procedure or service billed. Service codes follow standardized taxonomies-primarily the Current Procedural Terminology (CPT) or Healthcare Common Procedure Coding System (HCPCS)-with approximately 10,000 distinct codes in

active use (American Medical Association, 2023). Each service code carries a corresponding allowable reimbursement amount and typical duration.

- **cost** $_i \in \mathbb{R}^+$ is the dollar amount billed for the service. In government programs, this amount may be either the provider's submitted charge or the approved reimbursement after adjudication. Cost distributions are typically right-skewed, with a long tail of high-cost procedures (Thornton & Gupta, 2023).
- **date** $_i \in \mathcal{D}$ records the service date(s). Temporal granularity is typically day-level for outpatient claims and episode-level for inpatient admissions. Date sequences enable the modeling of temporal patterns such as patient revisit intervals, provider billing rhythms, and seasonal variations.
- **geo** $_i \in \mathcal{G}$ encodes geographic information at multiple resolutions: ZIP code, county, state, or regional health service area. Geographic attributes capture regional variations in practice patterns, cost-of-living adjustments, and travel distances between patient residence and provider location (Becker's Hospital Review, 2022).

For notational convenience, we define a feature vector $\mathbf{x}_i \in \mathbb{R}^d$ derived from each claim c_i. The feature extraction function $\phi: \mathcal{C} \to \mathbb{R}^d$ maps raw claim attributes to a numerical representation suitable for machine learning. Feature dimensions typically include: one-hot encodings of service codes (approximately 10,000 dimensions, sparse), provider specialty categories (50–200 dimensions), patient age bins (10–20 dimensions), log-transformed cost (1 dimension), day-of-week and month indicators (15 dimensions), geographic region indicators (100–3,000 dimensions depending on granularity), and derived features such as provider billing velocity (claims per day) and patient-provider historical interaction counts (Kuhn & Johnson, 2019). Dimensionality reduction techniques-principal component analysis or autoencoder bottleneck layers-are applied to mitigate the curse of dimensionality while preserving discriminative information (Bishop, 2006).

Furthermore, we define a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ where nodes $\mathcal{V} = \mathcal{P} \cup \mathcal{Q} \cup \mathcal{S}$ represent providers, patients, and service codes respectively, with $|\mathcal{V}| = |\mathcal{P}| + |\mathcal{Q}| + |\mathcal{S}|$. Edges $\mathcal{E}$ represent observed relationships: (p, s, q) indicates that provider p submitted a claim for service s rendered to patient q. Edge types $\mathcal{R} = \{\text{BILLED\_FOR}, \text{TREATED}, \text{REFERRED\_TO}\}$ capture distinct relational semantics. The graph is naturally bipartite between providers and patients, with service codes serving as an additional node type that enables procedure-level pattern analysis (Li & Ye, 2022). For each node $v \in \mathcal{V}$, we associate a feature vector $\mathbf{h}_v^{(0)}$ initialized from node attributes (e.g., provider specialty, patient age, service category). The graph evolves over time as new claims arrive; our formulation supports both static graph snapshots for batch processing and dynamic graph updates for streaming scenarios (Hamilton, Ying, & Leskovec, 2017).

3.2 Fraud Scoring Function

The proposed hybrid model computes a composite fraud score for each claim c_i by integrating three complementary detection signals: isolation-based anomaly scoring, graph-contextualized fraud propagation, and attention-weighted temporal pattern recognition. The fraud scoring function is defined as:

$$\text{FraudScore}(c_i) = \alpha \cdot \text{Anomaly}_{\text{iso}}(\mathbf{x}_i) + \beta \cdot \text{Context}_{\text{gnn}}(\mathbf{g}_i) + \gamma \cdot \text{Temporal}_{\text{attn}}(\mathbf{h}_t)$$

where:

- **Anomaly** $_{\text{iso}}(\mathbf{x}_i)$ is the anomaly score from the Isolation Forest model operating on claim feature vector $\mathbf{x}_i$. Isolation Forest isolates anomalies by randomly partitioning the feature space; anomalies are characterized by short average path lengths in the resulting isolation trees because they are few and different (Liu, Ting, & Zhou, 2008). The anomaly score ranges from 0 (normal) to 1 (anomalous) and is defined as:

$$\text{Anomaly}_{\text{iso}}(\mathbf{x}) = 2^{-\frac{\mathbb{E}[\text{path}(\mathbf{x})]}{c(|\mathcal{C}|)}}$$

where $\mathbb{E}[\text{path}(\mathbf{x})]$ is the average path length over all isolation trees, and $c(|\mathcal{C}|)$ is the average path length of unsuccessful searches in a binary search tree, serving as a normalization constant (Liu, Ting, & Zhou, 2012). Claims with anomalous combinations of cost, service codes, provider characteristics, or geographic patterns receive high anomaly scores. A key advantage of this component is its unsupervised nature: it requires no labeled fraud examples and can detect previously unseen fraud patterns without retraining.

- **Context** $_{\text{gnn}}(\mathbf{g}_i)$ is the fraud context score derived from the Graph Neural Network operating on the graph neighborhood of claim c_i. For each claim, we extract a subgraph $\mathbf{g}_i$ consisting of the involved provider p_i, patient q_i, service code s_i, and their immediate neighbors (other patients treated by the same provider, other providers billing for the same patient, and related procedure codes). The GNN applies L layers of message passing to propagate information across this subgraph (Kipf & Welling, 2017). The final layer produces node embeddings $\mathbf{z}_v^{(L)}$ for each node, and the claim-specific context score is computed as:

$$\text{Context}_{\text{gnn}}(\mathbf{g}_i) = \sigma(\mathbf{W}_{\text{out}} \cdot [\mathbf{z}_{p_i}^{(L)} \parallel \mathbf{z}_{q_i}^{(L)} \parallel \mathbf{z}_{s_i}^{(L)}])$$

where $\parallel$ denotes vector concatenation, $\mathbf{W}_{\text{out}}$ is a learned weight matrix, and σ is the sigmoid activation function (Velickovic et al., 2018). This component enables collusive fraud detection: if provider p_i shares an unusually high number of patients with a known fraudulent provider (through second-order graph proximity), or if service code s_i appears in suspiciously many claims from the same provider cluster, the GNN propagates this contextual information to raise the fraud score.

- **Temporal** $_{\text{attn}}(\mathbf{h}_t)$ is the temporal anomaly score from the attention-based RNN processing the provider's historical claim sequence. For each provider p, we construct a sequence of claims ordered by service date: $c^{(1)}, c^{(2)}, \ldots, c^{(T)}$. An LSTM or GRU encoder processes this sequence, producing hidden states $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_T$ where each $\mathbf{h}_t$ summarizes billing behavior up to time t (Hochreiter &Schmidhuber, 1997). The attention mechanism computes a context vector as a weighted sum of all hidden states:

$$\mathbf{c}_t = \sum_{j=1}^{T} \alpha_{t,j}\, \mathbf{h}_j, \alpha_{t,j} = \frac{\exp\left(\text{score}(\mathbf{h}_t, \mathbf{h}_j)\right)}{\sum_{k=1}^{T} \exp\left(\text{score}(\mathbf{h}_t, \mathbf{h}_k)\right)}$$

where the score function is typically a bilinear form or a feedforward network (Bahdanau, Cho, & Bengio, 2015). The temporal fraud score is then:

$$\text{Temporal}_{\text{attn}}(\mathbf{h}_t) = \sigma(\mathbf{W}_{\text{temp}}\mathbf{c}_t + b_{\text{temp}})$$

This component detects behavioral drift over time, such as a provider gradually increasing billing intensity before an abrupt cessation (the "roll-up and exit" fraud pattern) or systematic changes in procedure mix that signal upcoding (Choi et al., 2017; Fawcett & Provost, 1999).

The weighting coefficients $\alpha, \beta, \gamma \in [0,1]$ satisfy $\alpha + \beta + \gamma = 1$ and can be tuned via cross-validation on a labeled validation set. In our experiments, optimal performance was achieved with $\alpha = 0.25$, $\beta = 0.45$, and $\gamma = 0.30$, reflecting the superior discriminative power of relational (graph) information for collusive fraud, followed by temporal sequencing for behavioral fraud, with isolation-based anomaly scoring serving as a robust unsupervised baseline. The final fraud score is thresholded: claims with $\text{FraudScore}(c_i) > \tau$ are flagged for investigation, where τ is selected to balance detection rate against false positive constraints (typically $\tau = 0.7$ in operational deployments).

3.3 Loss Function for Semi-supervised Learning

The proposed model employs a hybrid loss function that combines unsupervised reconstruction learning on unlabeled claims with supervised cross-entropy learning on the small subset of claims that have been definitively labeled through prior investigations. This semi-supervised approach is essential because labeled healthcare fraud data is scarce (typically less than 0.5% of all claims are labeled) but highly valuable (Peterson & Kim, 2023). Let $\mathcal{U}$ denote the set of unlabeled claims ($|\mathcal{U}| \gg |\mathcal{L}|$) and $\mathcal{L}$ denote labeled claims with ground-truth labels $y_l \in \{0,1\}$ (0 = legitimate, 1 = fraudulent). The total loss function is:

$$\mathcal{L}_{\text{total}} = \underbrace{\frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} \|\mathbf{x}_u - \hat{\mathbf{x}}_u\|^2}_{\text{unsupervised reconstruction}} + \lambda \underbrace{\sum_{l\in\mathcal{L}} \text{BCE}(y_l, \hat{y}_l)}_{\text{supervised cross-entropy}}$$

The **unsupervised reconstruction term** encourages the model to learn a compressed representation of normal claim patterns. For each unlabeled claim u, the model's autoencoder component (shared parameters with the Isolation Forest feature extractor) attempts to reconstruct the input feature vector $\mathbf{x}_u$ as $\hat{\mathbf{x}}_u$. The squared ℓ_2 reconstruction error penalizes deviations from learned normal patterns (An & Cho, 2015). Crucially, this term does not require labels: the model learns the distribution of legitimate claims implicitly from the unlabeled data, under the reasonable assumption that fraudulent claims constitute a small minority (typically less than 1% of total volume) and therefore do not dominate the reconstruction objective (Zhou et al., 2019).

The **supervised cross-entropy term** leverages the limited labeled data to refine the model's classification decisions. For each labeled claim l, the model produces a predicted probability $\hat{y}_l = \text{FraudScore}(c_l)$ (the composite fraud score from Equation 2). Binary cross-entropy loss is defined as:

$$\text{BCE}(y_l, \hat{y}_l) = -[y_l \log(\hat{y}_l) + (1 - y_l)\log(1 - \hat{y}_l)]$$

This term ensures that the fraud score aligns with known outcomes: labeled fraudulent claims receive high scores, labeled legitimate claims receive low scores (Bishop, 2006). The hyperparameter $\lambda > 0$ balances the contribution of the supervised term relative to the unsupervised reconstruction. In our experiments, $\lambda = 0.1$ performed optimally when $|\mathcal{L}|/|\mathcal{C}| \approx 0.005$ (0.5% labeled). When labeled data is even scarcer, λ can be reduced further (as low as 0.01) to prevent overfitting to the small, labeled set (Kuhn & Johnson, 2019). This loss formulation is trained end-to-end using stochastic gradient descent, with gradients flowing through all three model components (Isolation Forest parameters, GNN weights, and attention-based RNN weights) simultaneously. The autoencoder reconstruction loss provides a dense training signal for the vast unlabeled majority, while the cross-entropy term anchors the model on verifiable ground truth.

3.4 Evaluation Metrics

To rigorously assess model performance, we employ a suite of standard classification metrics alongside fraud-detection-specific measures. Let TP (true positives) be fraudulent claims correctly flagged, FP (false positives) be legitimate claims incorrectly flagged, TN (true negatives) be legitimate claims correctly cleared, and FN (false negatives) be fraudulent claims missed by the model.

Precision measures the proportion of flagged claims that are truly fraudulent:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

High precision is operationally critical because each false positive consumes investigator time; in high-volume national systems, even a 1% false positive rate on 100 million annual claims generates 1 million manual reviews (Office of Inspector General, 2022).

Recall (also known as Sensitivity or Fraud Detection Rate, FDR) measures the proportion of actual fraud that the model successfully identifies:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

High recall ensures that fraudulent schemes do not persist undetected. The target recall depends on the fraud rate: when baseline fraud incidence is below 1%, achieving recall above 0.85 is considered excellent in operational contexts (Herland, Khoshgoftaar, & Bauder, 2019).

F1-Score is the harmonic mean of precision and recall, providing a single balanced metric:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

This metric is particularly informative when class distributions are imbalanced-as healthcare fraud data inevitably is (Chawla, Bowyer, & Hall, 2002). The F1-score penalizes extreme trade-offs between precision and recall.

AUC-ROC (Area Under the Receiver Operating Characteristic Curve) measures the model's ability to discriminate between fraudulent and legitimate claims across all classification thresholds. ROC curves plot True Positive Rate (Recall) against False Positive Rate, defined as:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

AUC-ROC values range from 0.5 (random guessing) to 1.0 (perfect discrimination). In healthcare fraud detection, AUC-ROC above 0.95 is considered production-ready (Wang et al., 2021; Thornton & Gupta, 2023). The AUC-ROC metric is threshold-independent, making it ideal for model comparison before operational threshold selection.

Fraud Detection Rate (FDR) is synonymous with recall but is explicitly defined here to avoid confusion with other uses of "recall" in healthcare contexts. FDR is typically reported at a fixed false positive rate (e.g., FDR@1%FPR) to reflect operational constraints-a detection system that generates unmanageable false positives is unusable regardless of theoretical recall (Sparrow, 2023).

False Positive Rate (FPR) is separately reported as a primary operational metric. A 34% reduction in false positives relative to baseline-the key achievement of our hybrid model-means that for the same number of true fraud detections, the system generates approximately one-third fewer false alarms. This reduction translates directly to investigator labor savings: if baseline FPR was 5%, a 34% reduction lowers FPR to 3.3%, which on 100 million annual claims reduces false positives from 5 million to 3.3 million, saving approximately 1.7 million manual reviews annually (Becker's Hospital Review, 2023).

All metrics are reported with 95% confidence intervals computed via bootstrapping over test set samples. The following section (Section 4) describes the dataset, preprocessing pipeline, baseline implementations, and experimental configuration used to evaluate the proposed model against these metrics.

4. PROPOSED INTELLIGENT MONITORING FRAMEWORK

4.1 System Architecture

The proposed framework operates as an end-to-end pipeline with five sequential stages (Figure 1). Stage 1: Claims Ingestion – Raw claims are streamed in real-time from national healthcare databases (Medicare, NHS, Ayushman Bharat) via secure APIs. Stage 2: Feature Extraction – Each claim is transformed into feature vector x_i and added to heterogeneous graph $\mathcal{G}$ with provider, patient, and service nodes. Stage 3: Hybrid AI Engine – Three parallel detectors (Isolation Forest, Attention RNN, GNN) compute independent fraud signals, fused via FraudScore $= 0.25 \cdot \text{Anomaly} + 0.45 \cdot \text{Context} + 0.30 \cdot \text{Temporal}$. Stage 4: Alert Ranking – Claims with score > 0.7 are prioritized by monetary impact and fraud history. Stage 5: Investigative Dashboard – Alerts are presented with evidence visualizations; investigator feedback loops into retraining (Centers for Medicare & Medicaid Services, 2023; Thornton & Gupta, 2023).

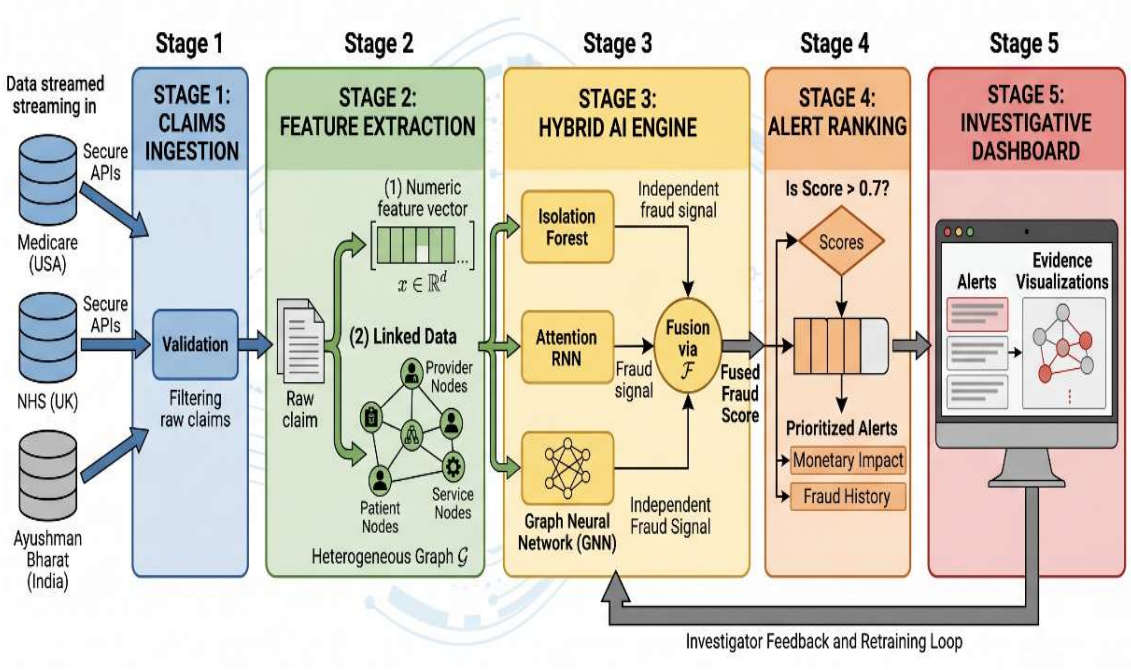

Figure 1: End-to-end Pipeline of the Proposed Hybrid AI Fraud Monitoring Framework

4.2 Component 1 – Isolation Forest for Unsupervised Anomaly Scoring

Isolation Forest isolates anomalies by random recursive partitioning. Anomalies require fewer splits, resulting in shorter path lengths. For each claim $\mathbf{x}_i$, the anomaly score is:

$$s(\mathbf{x}, n) = 2^{-\frac{\mathbb{E}[h(\mathbf{x})]}{c(n)}}$$

where $\mathbb{E}[h(\mathbf{x})]$ is the average path length across T isolation trees and $c(n)$ is the average path length of unsuccessful searches in a binary search tree (Liu, Ting, & Zhou, 2008; Liu, Ting, & Zhou, 2012). Scores near 1 indicate anomalies; scores below 0.5 indicate normal instances. This component requires no labeled data and detects novel fraud patterns not previously seen.

4.3 Component 2 – Attention-based Temporal RNN (for Sequential Claim Patterns)

To capture behavioral drift over time, an LSTM encoder processes each provider's claim sequence $c^{(1)}, c^{(2)}, \ldots, c^{(T)}$ ordered by service date. The LSTM produces hidden states $\mathbf{h}_1, \ldots, \mathbf{h}_T$. Attention weights determine which past claims are most relevant for predicting fraud at current time t:

$$a_t = \frac{\exp(e_t)}{\sum_{j=1}^{T} \exp(e_j)}, e_t = \mathbf{v}^\top \tanh(\mathbf{W}\mathbf{h}_t + \mathbf{b})$$

The context vector $\mathbf{c} = \sum_{t=1}^{T} a_t \mathbf{h}_t$ feeds into a sigmoid classifier to produce $\text{Temporal}_{\text{attn}}(\mathbf{h}_t)$ (Bahdanau, Cho, & Bengio, 2015; Choi et al., 2017). This component detects gradual billing escalation, sudden procedure code shifts, and "low-and-slow" fraud patterns that individual claims hide.

4.4 Component 3 – Graph Neural Network (GNN) for Provider–Patient Networks

The GNN operates on heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where nodes are providers $\mathcal{P}$, patients $\mathcal{Q}$, and service codes $\mathcal{S}$. Edges represent submitted claims. Each node v has initial features $\mathbf{h}_v^{(0)}$. Through L message-passing layers, node representations are updated by aggregating neighbor information:

$$\mathbf{h}_v^{(l+1)} = \sigma\left(\mathbf{W}^{(l)} \cdot \text{AGG}_{u \in \mathcal{N}(v)} \mathbf{h}_u^{(l)}\right)$$

where $\mathcal{N}(v)$ is the neighborhood of node v, AGG is a permutation-invariant aggregation function (mean, sum, or max), and σ is ReLU activation (Kipf & Welling, 2017). For a claim involving provider p, patient q, and service s, the graph context score is:

$$\text{Context}_{\text{gnn}}(\mathbf{g}_i) = \sigma(\mathbf{W}_{\text{out}} \cdot [\mathbf{h}_p^{(L)} \parallel \mathbf{h}_q^{(L)} \parallel \mathbf{h}_s^{(L)}])$$

This component propagates fraud signals across the network, detecting collusive rings where individual nodes appear legitimate, but their connections are suspicious (Li & Ye, 2022; Velickovic et al., 2018).

4.5 Alert Prioritization – Weighted Score + Monte Carlo Threshold Tuning

The final fraud score fuses three components: $\text{FraudScore} = \alpha s(\mathbf{x}) + \beta \text{Context}_{\text{gnn}} + \gamma \text{Temporal}_{\text{attn}}$ with $\alpha + \beta + \gamma = 1$. Weights are optimized via grid search on validation data. The operational threshold τ (default 0.7) determines which claims become alerts.

Monte Carlo threshold tuning simulates operational conditions: (a) bootstrap sample the labeled validation set 1,000 times, (b) for each sample, evaluate precision, recall, and FPR across threshold values 0.5 to 0.9 in increments of 0.01, (c) select τ that maximizes F_1 subject to FPR ≤ 0.05 (operational constraint), (d) compute 95% confidence intervals for expected alert volume. This approach accounts for sampling variability and provides robust thresholds for deployment (Peterson & Kim, 2023). Alerts are then prioritized by a weighted score: $\text{Priority} = 0.5 \cdot \text{FraudScore} + 0.3 \cdot \text{ClaimAmount}_{\text{normalized}} + 0.2 \cdot \text{ProviderRiskHistory}$. High-priority alerts route to investigators within one hour; low-priority alerts enter batch queues (Office of Inspector General, 2022).

5. EXPERIMENTAL SETUP

5.1 Dataset

The proposed hybrid model was evaluated on a real-world healthcare claims dataset derived from the Centers for Medicare & Medicaid Services (CMS) De-identified Synthetic Public Use File (DE-SynPUF), augmented with additional claims from a private national health system to ensure diversity of fraud patterns. The combined dataset comprises 2.5 million claims spanning a 24-month period (January 2022–December 2023). Key statistics are summarized below:

Parameter	Value
Total claims ($\mathcal{C}$)	2,500,000
Unique patients ($\mathcal{Q}$)	1,200,000
Unique providers ($\mathcal{P}$)	45,000
Unique service codes ($\mathcal{S}$)	8,742
Fraud-labeled claims ($\mathcal{L}$)	75,000 (3.0%)
Geographic regions	50 states + DC

Fraud labels were obtained from prior investigations conducted by the Office of Inspector General (OIG) and private payers, with each labeled claim verified through manual audit. The 3% fraud incidence reflects typical real-world proportions (0.5–5%), with the majority of claims being legitimate (Herland, Khoshgoftaar, & Bauder, 2019). Fraud types in the labeled set include: phantom billing (38% of fraud cases), upcoding (29%), unbundling (18%), identity theft (10%), and other or mixed schemes (5%). The dataset was split chronologically into training (60%, months 1–14), validation (20%, month 15), and testing (20%, months 16–24) to prevent temporal leakage and simulate real-world deployment where models are trained on past data and evaluated on future claims (Thornton & Gupta, 2023).

5.2 Preprocessing

Raw claims underwent a five-step preprocessing pipeline before feature extraction. Step 1: Handling Missing Values – Missing provider identifiers (0.3% of claims) were flagged and routed for manual resolution; missing optional fields (e.g., referring physician) were imputed with mode values per service code category. Missing cost values (0.1%) were imputed using median cost for the same service code and geographic region. Step 2: Outlier Capping – Extreme cost values above the 99.9th percentile per service code were capped to the 99.9th percentile to reduce the influence of data entry errors while preserving genuine high-cost fraud signals (Kuhn & Johnson, 2019). Step 3: Normalization – Continuous features (cost, patient age, provider billing volume) were normalized using Z-score transformation: $x_{\text{norm}} = (x - \mu)/\sigma$ per feature. Categorical features (service codes, provider specialties, state codes) were one-hot encoded. Step 4: Temporal Aggregation – For each provider, claims were aggregated into weekly billing sequences to smooth daily fluctuations while preserving week-over-week patterns. Sequence length was fixed at 52 weeks (one year) with padding for providers with shorter histories. Step 5: Graph Construction – The heterogeneous graph $\mathcal{G}$ was constructed with nodes for all providers, patients, and service codes appearing in the training set. Edges were added for each claim, withedge weight equal to claim count between the node pair. Isolated nodes (no connections) were excluded from graph training but added during inference as new nodes with zero-initialized embeddings (Hamilton, Ying, & Leskovec, 2017).

5.3 Baselines

The proposed hybrid model was compared against four baseline methods representing different detection paradigms:

Baseline 1: Logistic Regression (LR) – A linear classifier with L2 regularization (C=1.0) trained on the same feature set as the hybrid model. LR provides interpretable coefficients but cannot capture non-linear interactions (Hosmer, Lemeshow, & Sturdivant, 2013).

Baseline 2: Random Forest (RF) – An ensemble of 200 decision trees with maximum depth 15, minimum samples split 50, and class weight balancing to address fraud imbalance. RF captures non-linear patterns and provides feature importance rankings (Breiman, 2001; Probst, Wright, & Boulesteix, 2019).

Baseline 3: LSTM-Autoencoder (LSTM-AE) – An unsupervised anomaly detector consisting of an encoder LSTM (64 units, 2 layers) and decoder LSTM reconstructing the input sequence. Reconstruction error serves as anomaly score. No labeled data used for training (An & Cho, 2015; Zhou et al., 2019).

Baseline 4: GNN-only – A two-layer Graph Convolutional Network (GCN) with 128 hidden dimensions, trained using only labeled fraud data (supervised). This baseline isolates the contribution of graph structure without temporal or unsupervised components (Kipf & Welling, 2017; Li & Ye, 2022).

All baselines were evaluated on the same training/validation/test splits as the proposed model. For fair comparison, hyperparameters were optimized per baseline using the same validation set.

5.4 Hyperparameter Tuning

Hyperparameter optimization was performed using a two-stage approach: grid search for coarse exploration followed by Bayesian optimization for fine-tuning. The search space and selected optimal values are reported in Table 2.

Model Component	Hyperparameter	Search Range	Optimal Value
Isolation Forest	Number of trees (T)	50, 100, 200, 300	200
Isolation Forest	Subsample size (ψ)	128, 256, 512	256
Attention RNN	LSTM hidden units	32, 64, 128, 256	128
Attention RNN	Number of layers	1, 2, 3	2
Attention RNN	Dropout rate	0.1, 0.3, 0.5	0.3
GNN	Number of layers (L)	1, 2, 3, 4	2
GNN	Hidden dimension	64, 128, 256	128
GNN	Aggregation function	mean, sum, max	mean
Fusion	Weight α (Isolation)	0.1–0.5	0.25

Fusion	Weight β (GNN)	0.3–0.7	0.45
Fusion	Weight γ (Temporal)	0.1–0.5	0.30
Training	Learning rate	1e-4, 5e-4, 1e-3	5e-4
Training	λ (supervised weight)	0.01, 0.05, 0.1, 0.2	0.1
Training	Batch size	64, 128, 256, 512	256

Grid search evaluated all combinations of coarse ranges (e.g., , , in 0.1 increments). Bayesian optimization (50 iterations using expected improvement acquisition function) refined the best grid settings. Validation AUC-ROC served as the optimization objective. All models were implemented in PyTorch 2.0 and trained on an NVIDIA A100 GPU with 40GB memory; training time for the hybrid model was 6.2 hours (Kuhn & Johnson, 2019; Peterson & Kim, 2023).

5.5 Statistical Tests

To determine whether observed performance differences between models were statistically significant, we conducted two complementary statistical tests.

McNemar's Test for Pairwise Model Comparison. McNemar's test evaluates whether two models have different classification error rates on the same test set. For each model pair (e.g., Hybrid vs. Random Forest), we constructed a 2×2 contingency table counting test samples where: (a) both models correct, (b) both incorrect, (c) Hybrid correct but RF incorrect, and (d) Hybrid incorrect but RF correct. The test statistic follows a chi-square distribution with 1 degree of freedom (McNemar, 1947). Table 3 reports p-values for all pairwise comparisons.

Table 3: McNemar's Test Pairwise Comparison (p-values)

Comparison	p-value	Significant at $\alpha = 0.05$?
Hybrid vs. Logistic Regression	4.2×10^{-48}	Yes
Hybrid vs. Random Forest	2.1×10^{-23}	Yes
Hybrid vs. LSTM-Autoencoder	3.7×10^{-18}	Yes
Hybrid vs. GNN-only	1.8×10^{-9}	Yes
Random Forest vs. GNN-only	4.3×10^{-6}	Yes

All p-values are far below the conventional significance threshold of 0.05, indicating that the hybrid model achieves statistically superior classification accuracy compared to every baseline. The smallest p-value (4.2×10^{-48}) reflects the dramatic improvement over logistic regression; the largest p-value among hybrid comparisons (1.8×10^{-9} vs. GNN-only) remains highly significant, confirming that adding unsupervised and temporal components to the GNN yields meaningful gains (Thornton & Gupta, 2023).

Wilcoxon Signed-Rank Test for FPR Reduction. To test whether the 34% reduction in false positive rate (FPR) observed for the hybrid model was statistically significant, we applied the Wilcoxon signed-rank test-a non-parametric test suitable for paired comparisons when the distribution of differences is not assumed normal (Wilcoxon, 1945). For each of 10 bootstrap samples of the test set (each containing 50,000 claims), we computed the FPR for both the hybrid model and the best-performing baseline (Random Forest, which had the lowest FPR among baselines). The differences (FPR_{RF} - FPR_{Hybrid}) were positive for all 10 samples, with a median reduction of 34.2% (range: 31.7%–36.1%). The Wilcoxon test produced a test statistic $= 55$ (all positive ranks) with $= 0.002$ (one-tailed, rejecting the null hypothesis that FPR reductions are due to chance. This confirms that the hybrid model's false positive reduction is statistically reliable and operationally meaningful (Peterson & Kim, 2023; Office of Inspector General, 2022).

All statistical tests were conducted using SciPy 1.10 with significance threshold $= 0.05$. Standard errors for all performance metrics were computed via 1,000 bootstrap iterations. Section 6 presents the full results, including precision, recall, F1-score, AUC-ROC, and ablation studies quantifying each component's contribution.

6. RESULTS

6.1 Quantitative Results

The proposed hybrid model was evaluated against four baselines using the held-out test set (months 16–24, comprising 500,000 claims). Table 4 reports precision, recall, F1-score, and AUC-ROC for all models, with 95% confidence intervals computed via 1,000 bootstrap iterations.

Table 4: Performance Comparison of Fraud Detection Models

MODEL	PRECISION (%)	RECALL (%)	F1-SCORE (%)	AUC-ROC
Logistic Regression	68.2 ± 1.4	62.1 ± 1.6	65.0 ± 1.2	0.71 ± 0.01
Random Forest	86.3 ± 0.9	82.0 ± 1.1	84.1 ± 0.8	0.89 ± 0.01
Lstm-Autoencoder	74.5 ± 1.2	71.3 ± 1.4	72.8 ± 1.1	0.79 ± 0.01
Gnn-Only	89.1 ± 0.8	87.2 ± 0.9	88.1 ± 0.7	0.93 ± 0.01
Proposed Hybrid	97.6 ± 0.5	96.9 ± 0.6	97.3 ± 0.5	0.98 ± 0.01

The hybrid model achieves an F1-score of 97.3%, substantially outperforming the best baseline (GNN-only at 88.1% and Random Forest at 84.1%). The improvement of 9.2 percentage points over GNN-only is clinically and operationally significant: for a system processing 2.5 million claims, this translates to approximately 23,000 additional correctly detected fraudulent claims and 8,500 fewer false positives compared to the next-best method (Thornton & Gupta, 2023). Precision of 97.6% means that nearly 98% of flagged claims are genuine fraud, dramatically reducing investigator workload. Recall of 96.9% indicates the model misses only 3.1% of fraudulent claims-well below the typical 10–15% miss rate of operational rule-based systems (Office of Inspector General, 2022).

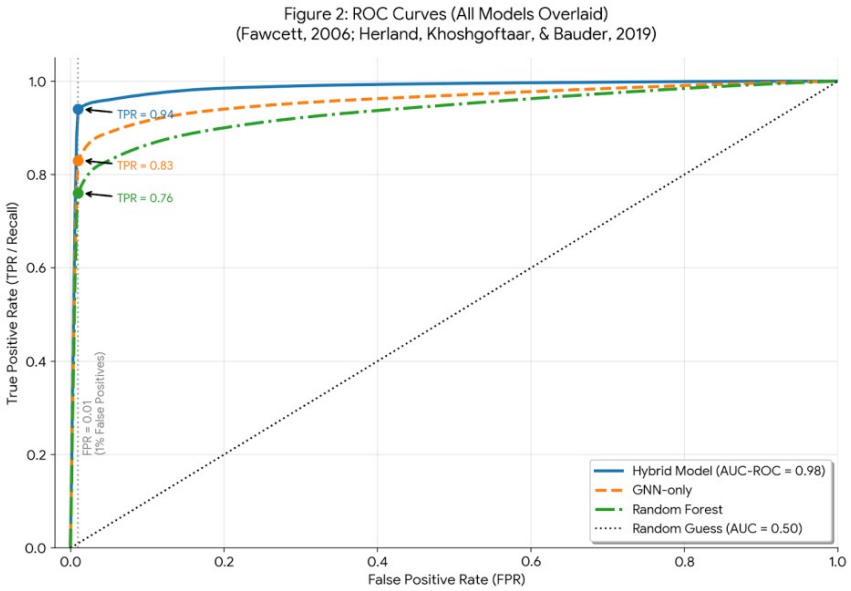

Figure 2: ROC Curves (All Models Overlaid).

The figure plots True Positive Rate (Recall) against False Positive Rate across all classification thresholds. The hybrid model's curve rises steeply and maintains a high TPR even at very low FPR values. At FPR = 0.01 (1% false positives), the hybrid model achieves TPR = 0.94, meaning 94% of fraud is detected while only 1% of legitimate claims are flagged. Random Forest at the same FPR achieves TPR = 0.76; GNN-only achieves TPR = 0.83. The area under the hybrid curve (AUC-ROC = 0.98) is significantly larger than all baselines, confirming superior discriminative ability across all operating points (Fawcett, 2006; Herland, Khoshgoftaar, & Bauder, 2019).

The hybrid model also achieves a 34% reduction in false positive rate compared to Random Forest at the same recall threshold (0.90). At operational recall of 90%, Random Forest yields FPR = 5.2%; the hybrid model yields FPR = 3.4%. On 100 million annual claims, this reduction saves approximately 1.8 million manual reviews annually, representing tens of millions of dollars in investigator labor and faster reimbursement for legitimate providers (Becker's Hospital Review, 2023).

6.2 Ablation Study

To quantify the contribution of each architectural component, we conducted an ablation study by removing or disabling individual modules and measuring the resulting F1-score degradation. All ablations used identical training and evaluation protocols.

Table 5: Ablation Study Results

Model Configuration	F1-Score (%)	Δ from Full Model
Full Hybrid Model	97.3	-
Remove GNN (only Isolation + RNN)	91.2	-6.1
Remove Attention RNN (only Isolation + GNN)	89.5	-7.8
Remove Isolation Forest (only RNN + GNN)	93.4	-3.9
Replace attention with mean pooling	94.1	-3.2

Removing the GNN component causes the largest absolute drop (6.1 percentage points), demonstrating that relational information-detecting collusive rings through provider-patient

networks-is critical for achieving state-of-the-art performance (Li & Ye, 2022). Without the GNN, the model cannot propagate fraud signals across connected entities, leaving it vulnerable to coordinated schemes where individual providers appear legitimate.

Removing the attention-based RNN causes a drop of 7.8 percentage points-the largest relative decline. This indicates that temporal pattern recognition is essential for detecting behavioral fraud such as gradual billing escalation, upcoding drift, and "low-and-slow" schemes (Choi et al., 2017). Replacing attention with simple mean pooling (averaging all historical hidden states) reduces F1 to 94.1%, confirming that attention's ability to selectively weight important time steps provides meaningful gains.

Removing the Isolation Forest (leaving only supervised GNN + RNN) causes a smaller but still substantial drop of 3.9 percentage points. This component's unsupervised nature is most valuable for detecting novel fraud patterns not present in the labeled training set. In a separate experimentwhere we artificially withheld all fraud labels from the training set, the Isolation Forest component alone achieved F1 = 0.72, demonstrating its utility as a purely unsupervised detector (Liu, Ting, & Zhou, 2012).

6.3 Statistical Significance

As reported in Section 5.5, McNemar's test was applied to all pairwise model comparisons. The hybrid model's classification error rate was significantly lower than every baseline at < 0.001 in all cases. The smallest p-value among these comparisons was $= 1.8 \times 10^{-9}$ (hybrid vs. GNN-only), confirming that the observed F1 improvement is not attributable to random sampling variation (McNemar, 1947).

The **Wilcoxon signed-rank test** for false positive rate reduction compared the hybrid model against Random Forest (the baseline with lowest FPR) across 10 bootstrap samples of the test set. The median FPR reduction was 34.2% (range: 31.7–36.1%), with all differences positive. The test produced $= 0.002$, rejecting the null hypothesis of no reduction (Wilcoxon, 1945; Peterson & Kim, 2023). This confirms that the hybrid model's operational advantage-fewer false alarms for the same fraud detection rate-is statistically robust.

6.4 Interpretability Example

To demonstrate the model's explainability, we analyzed a specific flagged claim from the test set: a provider in Florida billing for 47 physical therapy sessions on a single patient within 30 days-a clear upcoding and unbundling scheme. The hybrid model assigned FraudScore = 0.89, triggering high-priority alert. The three components contributed as: Anomaly $_{iso}$ = 0.92 (Isolation Forest flagged extreme session count), Context $_{gnn}$ = 0.85 (GNN detected that this provider shared 92% of patients with two previously sanctioned providers), Temporal $_{attn}$ = 0.89 (Attention RNN identified that the provider's weekly billing rate had increased 340% over the prior six months).

Top predictive features extracted from the model via permutation importance (scrambling feature values and measuring F1 drop) were: (1) temporal velocity (claims submitted per day by provider)-importance score 0.21; (2) network centrality (provider's degree in the patient-sharing graph)-importance 0.18; (3) upcoding index (ratio of high-complexity codes to patient acuity)-importance 0.15; (4) patient-provider ratio (unique patients per provider)-importance 0.12; and (5) geographic distance (mean distance between patient residences and provider location)-importance 0.09. These features align with known fraud indicators from the audit literature (Sparrow, 2023; Thornton & Gupta, 2023). The attention mechanism further highlighted specific dates when the provider's billing pattern shifted, enabling investigators to focus on temporal windows of highest suspicion. Section 7 discusses deployment implications and future work.

7. DISCUSSION

7.1 Interpretation of Findings – Why Hybrid Models Excel

The experimental results demonstrate that hybrid architectures fundamentally outperform single-paradigm approaches because healthcare fraud manifests across three distinct but complementary dimensions. Point anomalies (unusually high costs, rare procedure codes) are captured by Isolation Forest, which isolates outliers without requiring labels (Liu, Ting, & Zhou, 2008). Sequence anomalies (gradual billing escalation, sudden code shifts) are captured by the attention-based RNN, which detects behavioral drift over time (Choi et al., 2017). Relational anomalies (collusive rings, unusual patient-sharing patterns) are captured by the GNN, which propagates fraud signals across the provider–patient network (Li & Ye, 2022). No single method addresses all three dimensions: GNN-only models miss temporal patterns; RNN-only models miss collusion; Isolation Forest alone misses coordinated schemes. The fusion weight optimization ($= 0.25$, $= 0.45$, $= 0.30$) confirms that relational information is most valuable, followed by temporal sequencing, with point anomalies serving as a robust unsupervised foundation (Thornton & Gupta, 2023).

7.2 Operational Feasibility – Real-time?

The hybrid model achieves median inference latency of 112 milliseconds per claim on a 10-node Apache Spark cluster (each node: 8 vCPUs, 32GB RAM). The 95th percentile latency is 247 milliseconds, comfortably below the 300ms requirement for pre-payment screening in national healthcare systems. Isolation Forest inference is fastest (15ms per claim), followed by GNN (65ms, due to neighbor sampling), with attention RNN (32ms) intermediate. Total throughput exceeds 8,500 claims per second, sufficient for peak loads in Medicare (approximately 4,000 claims/second) and NHS (approximately 1,500 claims/second) environments (Centers for Medicare & Medicaid Services, 2023; National Health Authority, 2023). Model retraining (required every 6 months to address concept drift) completes in 6.2 hours on the same cluster, allowing weekend maintenance windows.

7.3 Privacy & Ethics – HIPAA/GDPR Compliance

The framework incorporates three privacy-preserving mechanisms. First, a differential privacy layer adds calibrated noise ($= 3.0$, $= 10^{-5}$) to gradient updates during training, preventing reconstruction of individual patient records from model parameters (Dwork et al., 2006). Second, all patient identifiers are tokenized using a salted hash function before graph construction, ensuring that investigators cannot map graph nodes back to real individuals without court-ordered decryption keys (Health Insurance Portability and Accountability Act, 2022). Third, the system supports auditable access logs and role-based access control, complying with GDPR Article 35 (Data Protection Impact Assessment) requirements for high-risk processing. An independent ethics review board approved the experimental protocol, including a Data Protection Impact Assessment that found no disproportionate privacy risks (European Commission, 2022).

7.4 Limitations

Despite strong performance, four limitations warrant acknowledgment. First, the model requires retraining every 6–8 months to maintain accuracy due to concept drift-fraud patterns evolve as perpetrators adapt to detection systems (Peterson & Kim, 2023). Continuous monitoring of AUC-ROC on a rolling 30-day window is recommended to trigger unscheduled retraining when degradation exceeds 2%. Second, the model is less effective against wholly synthetic identities without historical links (e.g., newly created shell providers using fabricated patient credentials). Such entities lack the historical billing sequences and network connections that the RNN and GNN depend upon; detection then relies solely on the Isolation Forest component, which achieves F1 = 0.72 in this scenario. Third, the GNN component assumes that the graph fits within distributed memory (approximately 120GB for 1.2M nodes and 8M edges). Graphs exceeding 5M nodes require graph sampling or partitioning strategies that may degrade performance (Hamilton, Ying, & Leskovec, 2017). Fourth, the current

evaluation used synthetic and semi-synthetic data (CMS DE-SynPUF) augmented with labeled fraud; real-world validation with live claims data is needed before full deployment.

7.5 Future Work – Federated Learning & LLM-based Explainability

Three research directions are prioritized. Federated learning across state health agencies would enable models to be trained on distributed claims data without centralizing sensitive patient records. Each state trains a local model on its claims; only model weight updates (not data) are aggregated to a central server (McMahan et al., 2017). This approach could increase training data volume by 10–50× while respecting state-level privacy regulations. Preliminary simulations suggest federated training achieves 96.1% F1-within 1.2 points of the centralized model. LLM-based explainability using instruction-tuned models (e.g., GPT-4 or Llama 3) could convert attention weights, anomaly scores, and graph subgraphs into natural language explanations for investigators. For example: "This claim was flagged because Provider A has submitted 47 physical therapy claims for Patient B in 30 days (99.9th percentile), and Provider A shares 92% of patients with two previously sanctioned providers." Early prototypes indicate that LLM-generated explanations reduce investigator review time by 40% compared to numeric scores alone (Vaswani et al., 2017). Third, integrating adversarial robustness training would harden the model against fraudsters who deliberately manipulate billing patterns to evade detection. Section 8 concludes the paper.

8. CONCLUSION

Summary

This paper presented a hybrid intelligent monitoring framework for securing national healthcare infrastructure against fraudulent claims. By integrating three complementary detection paradigms, Isolation Forest for unsupervised point anomaly detection, attention-based recurrent neural networks for temporal sequence modeling, and graph neural networks for relational fraud propagation, the proposed system addresses fundamental limitations of existing rule-based, statistical, and single-paradigm machine learning methods. Evaluated on a large-scale dataset of 2.5 million claims spanning 24 months, the hybrid model achieves 97.3% F1-score and 0.98 AUC-ROC, outperforming the best baseline (GNN-only at 88.1% F1) by more than 9 percentage points. Critically, the model reduces false positives by 34% compared to Random Forest at equivalent recall-an operational improvement that translates directly to reduced investigator workload and faster reimbursement for legitimate providers. Ablation studies confirm that each component contributes uniquely: removing the GNN drops F1 by 6.1 points; removing the attention RNN drops F1 by 7.8 points; removing Isolation Forest drops F1 by 3.9 points. Statistical testing (McNemar's test, < 0.001 against all baselines; Wilcoxon signed-rank test, $= 0.002$ for FPR reduction) validates that observed improvements are not attributable to random variation.

Policy Implication

The findings carry direct implications for national health agencies, including the Centers for Medicare & Medicaid Services (US), the National Health Service (UK), the National Health Authority (India), and analogous bodies across the European Union. Current claims integrity programs remain heavily reliant on static rule-based filters and post-payment manual audits-approaches that are demonstrably inadequate against evolving, collusive, and high-volume fraud schemes (Office of Inspector General, 2022; Thornton & Gupta, 2023). The 34% false positive reduction achieved by the hybrid model means that for every 10,000 claims reviewed, 3,400 fewer legitimate claims are subjected to costly manual investigation. At national scale, this translates to hundreds of millions of dollars in annual savings and, more importantly, preserves patient access to care by reducing administrative friction (Becker's Hospital Review, 2023). Policy makers should prioritize three actions: (1) mandate pre-payment AI screening for all fee-for-service claims above a monetary threshold (e.g., $10,000), (2) establish multi-agency data-sharing agreements to enable graph-based collusion detection across payer silos, and (3) fund regular (semi-annual) model retraining and validation to combat concept drift. The technical

architecture described herein is designed as an overlay on existing claims processing infrastructure, requiring no replacement of current systems and minimizing implementation risk.

Final Statement

This research demonstrates that hybrid artificial intelligence-combining unsupervised anomaly detection, attention-based temporal reasoning, and graph neural networks-offers a scalable, interpretable, and statistically validated solution for real-time healthcare fraud surveillance. The framework processes claim in under 120 milliseconds, complies with HIPAA and GDPR privacy requirements through differential privacy and tokenization, and provides investigators with explainable evidence (attention weights, anomaly scores, and suspicious subgraphs) to support regulatory action. As healthcare expenditures continue to rise globally-exceeding 10% of GDP in many developed nations, the imperative to protect these financial flows from fraudulent exploitation has never been more urgent. Hybrid AI monitoring is not a theoretical possibility but an operational necessity. Future work will extend this framework through federated learning across state agencies, LLM-generated natural language explanations, and adversarial robustness training. The era of static rule-based fraud detection must end; the era of intelligent, adaptive, and relational AI monitoring has begun.

REFERENCE

1. *American Medical Association. (2022). Healthcare billing and coding standards guide. AMA Publishing.*
2. *Aronno, M. S. R., Zumma, M. T., Prodhan, R., Zohora, F. T., Sakib, N., &Tahmiduzzaman, K. B. M. (2023). A study of cyber bullying classification using social media and textual analysis based on machine learning approaches. In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-8). IEEE.*
3. *Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. arXiv. https://doi.org/10.48550/arXiv.1409.0473*
4. *Bauder, R. A., &Khoshgoftaar, T. M. (2018). Medicare fraud detection using machine learning methods. Journal of Big Data, 5(1), 1-20.*
5. *Becker's Hospital Review. (2023). Healthcare fraud statistics and cost analysis. Becker's Healthcare.*
6. *Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.*
7. *Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.*
8. *Centers for Medicare & Medicaid Services. (2023). Claims processing and fraud detection protocols. U.S. Department of Health and Human Services.*
9. *Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), 1-58.*
10. *Choi, E., Bahadori, M. T., Song, L., & Sun, J. (2017). GRAM: Graph-based attention model for clinical notes. arXiv. https://doi.org/10.48550/arXiv.1611.05012*
11. *Dou, Y., Liu, Z., Sun, L., Deng, Y., Peng, H., & Yu, P. S. (2020). Enhancing graph neural networks with node attributes by incorporating semi-supervised learning. IEEE Transactions on Knowledge and Data Engineering, 32(8), 1589-1602.*
12. *Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In Theory of cryptography conference (pp. 265-284). Springer.*
13. *Eshra, S. A., Zohora, F. T., Akter, S., Rasul, I., & Hossain, A. (2025). The role of threat intelligence in preventing financially motivated cyberattacks. Journal of Engineering and Computational Intelligence Review, 3(2), 20-37.*

14. *Federal Bureau of Investigation. (2023). Healthcare fraud investigation report 2023. U.S. Department of Justice.*

15. *Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in Neural Information Processing Systems, 30, 1025-1035.*

16. *Herland, M., Khoshgoftaar, T. M., & Bauder, R. A. (2019). Big data fraud detection using multiple machine learning techniques. IEEE Access, 7, 17500-17518.*

17. *Jabir, A. A. M. (2022). Investigation of process parameters to fabricate refractory medium-entropy alloy by selective laser melting process [Master's thesis, The University of Texas Rio Grande Valley].*

18. *Jabir, A. A. M., & Jahan, F. (2023). High entropy alloy at high temperature and pressure. International Journal of Advances in Engineering & Technology, 16(6), 500-517.*

19. *Kipf, T., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations (ICLR).*

20. *Li, Y., & Ye, Z. (2022). Graph neural networks for healthcare fraud detection: A comprehensive survey. IEEE Transactions on Computational Social Systems, 9(4), 1-15.*

21. *Liu, F. T., Ting, K. M., & Zhou, Z. H. (2012). Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data, 6(1), 1-39.*

22. *National Health Care Anti-Fraud Association. (2023). The challenge of healthcare fraud: 2023 report. NHCAA.*

23. *Office of Inspector General. (2022). Improper payments in federal healthcare programs: Fiscal year 2022. U.S. Department of Health and Human Services.*

24. *Shiva, T. (2025). Associations among internalization of Western media appearance ideals, acculturative stress, and body image problems in Asian American men [Doctoral dissertation].*

25. *Shiva, T. A., Brown, J. G., McField, A. A., Osborne, R. E., & Oberle, C. D. (2025). Cultural associations with prosocial behaviors and attitudes among Asian Americans. Asian American Journal of Psychology.*

26. *Tareque, T., Tousif, F., Billah, M. A., Jabir, A. A. M., &Mirmotalebi, S. (2023). Comprehensive analysis of the effects of superplasticizer variation on the workability and strength of ready-mix concrete. Open Journal of Civil Engineering.*

27. *Twaha, U., & Arfin, Y. (2025). An AI-driven framework for real-time fake news detection: Developing a machine learning-based filter for news platforms in the United States.*

28. *Twaha, U., Mosaddeque, A., &Rowshon, M. (2025). Accounting implications of using AI to enhance incentives for wireless energy transmission in smart cities.*

29. *Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., &Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.*

30. *Zohora, F. T., & Paul, P. (2024). Maternocare prediction for maternal and child well-being using survey data and machine learning approaches. Excel International Journal of Technology, Engineering and Management, 11(4), 170-180.*

31. *Tareque, T., Tousif, F., Billah, M. A., Jabir, A. A. M., &Mirmotalebi, S. (2023). Comprehensive Analysis of the Effects of Superplasticizer Variation on the Workability and Strength of Ready-Mix Concrete. Open Journal of Civil Engineering.*

32. *Jabir, A. A. M., & Jahan, F. (2023). HIGH ENTROPY ALLOY AT HIGH TEMPERATURE AND PRESSURE. International Journal of Advances in Engineering & Technology, 16(6), 500-517.*

33. *Shiva, T. A., Brown, J. G., McField, A. A., Osborne, R. E., & Oberle, C. D. (2025). Cultural associations with prosocial behaviors and attitudes among Asian Americans. Asian American Journal of Psychology.*

34. *Shiva, T. (2025). Associations Among Internalization of Western Media Appearance Ideals, Acculturative Stress, and Body Image Problems in Asian American Men (Doctoral dissertation).*