Graph-Based Detection Of Anomalous Health Insurance Claims Using Transformer-Augmented Embeddings

Dhanrithii D
Computer Science and Engineering
Sri Venkateswara College of Engineering
Sriperumbudur, India

Dr Suresh Kumar M
Computer Science and Engineering
Sri Venkateswara College of Engineering
Sriperumbudur, India

Abstract—Health insurance fraud results in significant financial losses and necessitates scalable, intelligent detection methods. This work presents a modular and unsupervised framework for fraud detection in healthcare claims by integrating semantic feature extraction, heterogeneous graph modeling, and anomaly detection. Structured claim metadata and unstructured clinical narratives are embedded using transformer-based models to capture rich contextual information. These embeddings are incorporated into a heterogeneous graph that connects patients, providers, and claims, enabling the use of graph neural networks to learn complex relational patterns indicative of fraudulent behavior. A graph-based representation is constructed using synthetic yet realistic healthcare data generated in standard clinical formats. Contextual node embeddings are learned, and clustering methods are applied to identify latent behavioral patterns. Unsupervised anomaly detection techniques, including tree-based and distance-based models, are employed to flag suspicious entities. A provider-level risk scoring mechanism is introduced to prioritize investigation efforts. This framework is designed to operate without reliance on labeled data, ensuring adaptability to evolving fraud strategies and generalizability across diverse claim scenarios. Experimental evaluation shows the system's effectiveness in uncovering subtle fraud signatures, highlighting its potential as a robust alternative to traditional rule-based approaches.

Keywords—healthcare fraud detection, graph neural networks, transformer models, anomaly detection, semantic embeddings

I. Introduction

Health insurance fraud has emerged as a critical challenge to healthcare systems globally, leading to financial losses estimated between $68 billion and $230 billion annually in the United States alone. With the growth of digital health records and the adoption of standardized formats such as FHIR (Fast Healthcare Interoperability Resources), there exists a significant opportunity to utilize advanced technologies for intelligent fraud detection. Traditional methods—including manual audits, rule-based systems, and supervised learning—face limitations in scalability and often fail to identify complex, evolving fraud patterns. Advanced approaches involving deep learning architectures such as Longformer have shown promise in extracting semantic insights from lengthy clinical texts, while Graph Neural Networks (GNNs) offer effective ways to model relationships between healthcare entities, enabling deeper understanding of systemic interactions.

The integration of Longformer-based semantic embeddings with heterogeneous GNNs allows for the detection of anomalous behaviour by jointly analysing unstructured clinical text and structured metadata. A healthcare knowledge graph is constructed using synthetic yet realistic claims data generated through the Synthea framework. The graph models claims, patients, and providers to uncover hidden behavioural clusters, detect anomalies, and identify provider-level fraud without relying on labelled data. This system addresses key challenges related to semantic-context integration and label scarcity, while enabling unsupervised, scalable fraud detection. Through this approach, a more adaptive and intelligent framework is established to ensure transparency and accountability within digital healthcare infrastructures.

II. BACKGROUND

Healthcare fraud detection has witnessed significant advancements with the rise of technologies such as machine learning, graph-based analytics, blockchain systems, and deep learning architectures. These developments aim to overcome the limitations of traditional rule-based systems by improving scalability, generalizability, and accuracy in detecting fraudulent behaviour. To contextualize the proposed ClaimNet framework, it is necessary to examine how existing paradigms compare across core dimensions such as graph-based modelling, unsupervised anomaly detection, and text integration.

Early approaches relied heavily on rule-based or statistical models, which used handcrafted heuristics and simple numerical thresholds to flag suspicious claims. While straightforward to implement, such systems lacked adaptability and struggled to capture complex or evolving fraud schemes. In contrast, recent graph-based methods model the relationships among healthcare entities. For instance, Hong et al. [1] and Lu et al. [11] proposed heterogeneous graph learning architectures that encode interactions between providers, claims, and policyholders using attention mechanisms and entity-aware embeddings. Meulemeester et al. [4] and Kennedy et al. [9] explored explainable anomaly detection pipelines, yet did not incorporate healthcare-specific graph structures. ClaimNet addresses this gap by integrating relational modelling with semantic text understanding to uncover latent fraud patterns.

Supervised machine learning remains a dominant paradigm. Du Preez et al. [3] highlighted the prevalence of supervised models in fraud detection pipelines, while also noting their dependence on labelled data. Hancock et al. [7] built ensemble-based systems combining feature selection with traditional classifiers, and Nabrawi et al. [12] assessed various supervised algorithms such as decision trees and SVMs. While these methods achieve high accuracy on structured inputs, they typically exclude unstructured clinical narratives and exhibit poor generalization to unseen fraud types. ClaimNet overcomes these limitations by operating in a fully unsupervised regime, leveraging both structured and unstructured data through semantic and graph-based representation learning.

Blockchain-enabled fraud prevention frameworks are gaining attention for their auditability and transparency. Amponsah et al. [2] and Jena et al. [8] developed smart contract-based solutions using Ethereum and Hyperledger Fabric, respectively. These systems improved traceability but lacked integration with learning-based anomaly detection models and struggled with scalability across diverse healthcare scenarios. In contrast, ClaimNet emphasizes analytical fraud detection by fusing transformer-based embeddings with graph neural reasoning and unsupervised outlier detection.

Text mining approaches have also been applied to health claims data. Gamaleldin et al. [5] used CNNs with Latent Dirichlet Allocation (LDA) to classify insurance risk, while Hamid et al. [6] utilized TF-IDF and rule-based heuristics to flag potential fraud. Tabassum et al. [15] explored anomaly detection in decentralized health systems, though their model lacked deep semantic contextualization. ClaimNet builds upon these efforts by using Longformer to encode long-form clinical narratives and provider notes, capturing richer semantic features critical for fraud profiling.

Finally, hybrid and anomaly-based models offer integrative perspectives. Li et al. [10] proposed a contrastive learning framework for anomaly detection in patient trajectory data, while Prakosa et al. [13] applied k-means clustering and regression to detect irregular billing. Rochner et al. [14] used unsupervised anomaly detection to flag implausible entries in cancer registries. Although these approaches demonstrate promise, they do not incorporate multimodal inputs or graph-aware learning. ClaimNet advances the field by providing an end-to-end, unsupervised pipeline that fuses semantic embeddings, heterogeneous graph modelling, and structural anomaly detection across multiple data types.

This review underscores the breadth of approaches in healthcare fraud detection and positions ClaimNet as a novel, multimodal framework that bridges critical gaps in unsupervised learning, semantic integration, and relational reasoning.

III. METHODOLOGIES

A. *Data Generation and Preprocessing*

The dataset used in this study was generated using Synthea [16], an open-source synthetic patient simulation tool that produces electronic health records in FHIR format. Structured tables were extracted from resources including Patient, Provider, Claim, Procedure, and Encounter. To simulate fraudulent activity, controlled anomalies were introduced such as upcoding, mismatched diagnosis-procedure pairs, and claim clustering from selected providers. These perturbations retained realistic distributions while embedding detectable fraudulent patterns.

Each claim included both structured metadata and unstructured textual elements such as diagnoses, procedures, and clinical notes. These were concatenated to form a unified document for each claim, as shown in (1):

$$d_i = \text{concat}\ (\text{diagnosis}_i, \text{procedure}_i, \text{notes}_i) \tag{1}$$

Text preprocessing involved lowercasing, punctuation removal, and token normalization. To refine the vocabulary, a frequency filter removed extremely rare and common tokens as defined in (2):

$$f(w_j) = \sum_{i=1}^{n} \text{count}(w_j, d_i) \tag{2}$$

Table I presents examples of raw and cleaned procedure texts.

TABLE I. SAMPLE PROCEDURE TEXT PREPROCESSING

Raw Procedure Text	Cleaned Text
['Amoxicillin 500 MG Oral Tablet']	amoxicillin mg oral tablet
['Acetaminophen 160 MG Chewable Tablet']	acetaminophen mg chewable tablet
['Encounter for symptom (procedure)', 'Otitis media (disorder)']	encounter for symptom procedure otitis media disorder

B. *Semantic Embedding Using Longformer*

To extract contextual embeddings, the cleaned claim texts were passed through Longformer [8], a transformer model designed for long sequences using a hybrid attention mechanism combining sliding window and global attention. Each document was truncated to 512 tokens and processed to yield contextual token embeddings, which were then mean-pooled into a fixed-size 768-dimensional vector per claim. This embedding captured semantic information from clinical narratives. Fig. 1 outlines the overall pipeline.

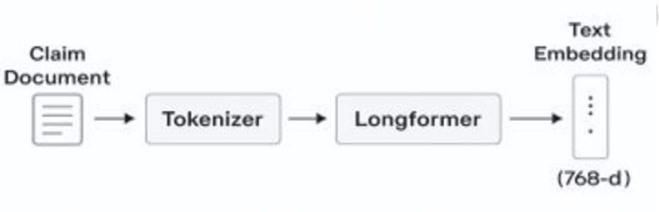

Fig. 1. Claim text embedding pipeline using Longformer

C. *Heterogeneous Graph Construction*

A heterogeneous graph was constructed using the Deep Graph Library (DGL) [17] to capture structural relationships among claims, patients, and providers. Nodes represented healthcare entities, and edges encoded typed relationships such as belongs_to (claim → patient) and processed_by (claim → provider). Reverse edges (has_claim, processed) enabled bidirectional message passing, consistent with graph-based fraud modelling strategies [1], [11].

Node features were defined as:

- Claim nodes: Six-dimensional metadata and 768-dimensional Longformer embeddings
- Patient and provider nodes: Statistical aggregates (e.g., visit frequency, average claim value)

The complete graph is shown in Fig. 2, and a zoomed subgraph in Fig. 3 illustrates real-world relationships.

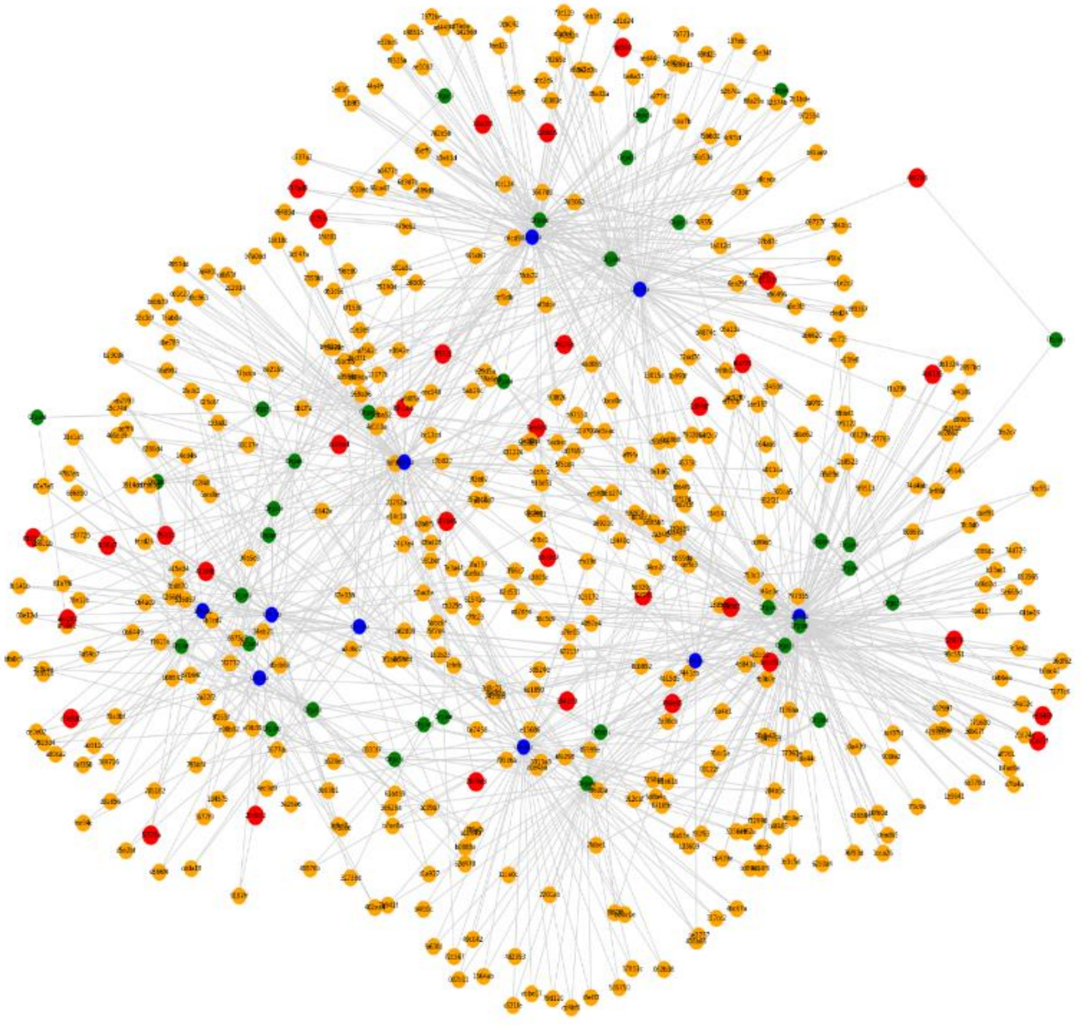

Fig. 2. Full graph of Claims, Patients and Providers with Fraud Highlighted

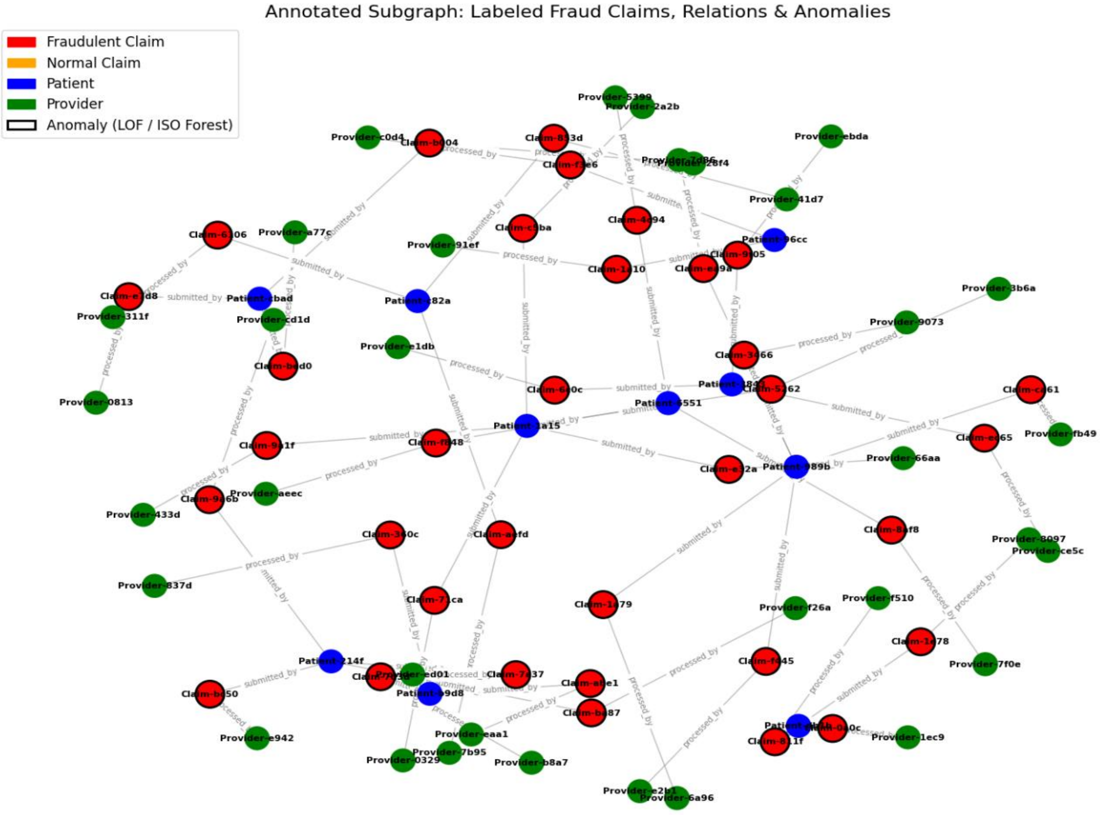

Fig. 3. An annotated subgraph of the claim-patient-provider network

Table II summarizes the number of nodes and dimensionality of features used for each entity type.

TABLE II. FEATURE MATRIX SUMMARY BY NODE TYPE

Node Type	**Number of Nodes**	**Feature Dimensionality**
Claim	502	6
Patient	446	2
Provider	503	3

D. *Heterogeneous GNN Architecture*

A two-layer heterogeneous Graph Neural Network was implemented using DGL [17], with separate GraphConv layers per edge type to facilitate relation-specific message passing. ReLU activation, dropout, and batch normalization were applied to improve stability. After two layers, claim nodes were projected into a 16-dimensional latent space. This design follows the multi-channel heterogeneous GNN structure proposed in [1], and supports learning complex interaction patterns relevant to fraud detection, as also emphasized in [3].

E. *Multi-Modal Embedding Fusion*

Each claim node was represented by a fused embedding combining:

- Structured metadata (6-dim)
- Longformer semantic vector (768-dim)
- GNN structural embedding (16-dim)

The result was a 790-dimensional feature vector per claim capturing structured, textual, and relational information. This fusion empowered downstream models to identify both anomalous claims and suspicious connectivity patterns.

F. *Anomaly Detection Techniques*

To detect potential fraud, two unsupervised anomaly detectors were applied on the fused embeddings:

1. Isolation Forest: Detects claims easily isolated by random partitions, indicating extreme or rare features [4].
2. Local Outlier Factor (LOF): Flags claims with significantly lower neighbourhood density, indicating local deviation [3], [6].

Claims flagged by both were marked as high-confidence anomalies. To assess fraud risk at the provider level, a risk score was computed based on the average anomaly scores of that provider's claims, as in (3):

$$\text{Risk}(p_k) = (\sum_{(}c \in C_{p)}\ \text{AnomalyScore}(c))/\ |C_p| \qquad (3)$$

where Cp is the set of claims submitted by provider pkp_kpk. This aligns with explainable and scalable fraud detection practices [7].

IV. RESULTS AND DISCUSSION

1) *Clustering and Embedding Analysis*

To assess whether the fused embeddings effectively differentiated claim behaviors, K-Means clustering was applied on the final 790-dimensional claim representations. The silhouette score indicated the optimal number of clusters to be k = 5, suggesting clear separation. These clusters reflected distinct behavioral trends, including differences in billing amounts, insurance coverage, and claim frequencies.

Dimensionality reduction via Uniform Manifold Approximation and Projection (UMAP) was employed for visualization. As shown in Fig. 4, anomalous claims predominantly appeared in loosely connected, peripheral areas of the 2D embedding space. This spatial separation suggests that fraudulent claims share latent semantic and relational patterns, effectively captured by the multimodal fusion approach.

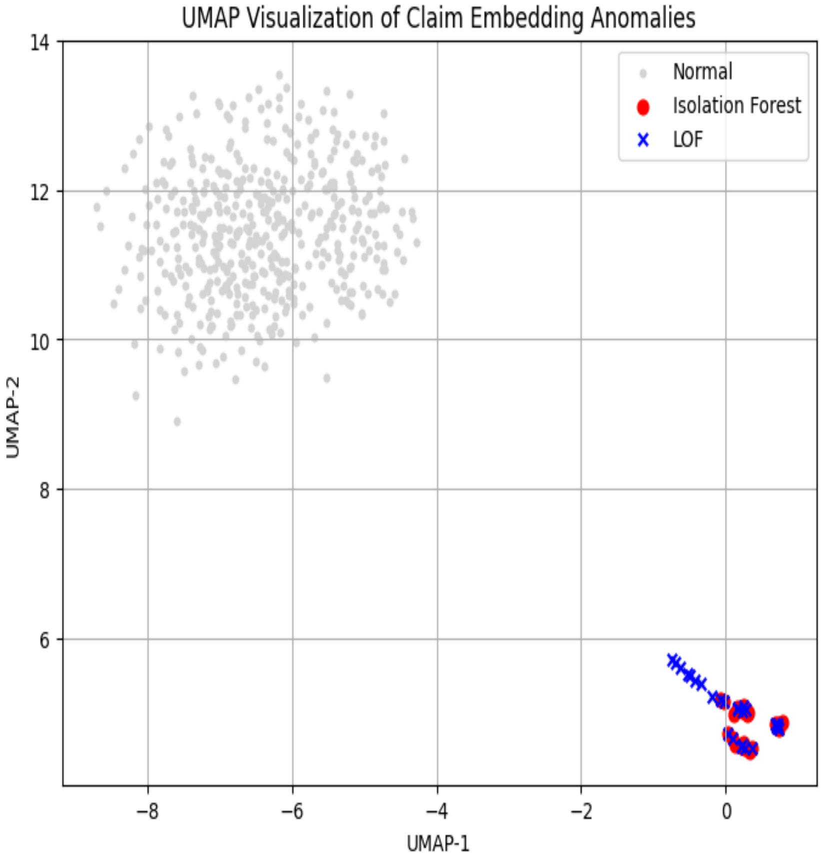

Fig. 4. UMAP Visualization of Claim Embedding Anomalies

2) *Anomaly Characteristics and Distribution*

To detect anomalous claims, both Isolation Forest and Local Outlier Factor (LOF) were applied independently to the fused embeddings. Each method identified 26 outliers, with 18 claims overlapping between both detectors. The Jaccard Index for the intersection was 0.53, indicating moderate agreement and supporting the utility of ensemble-based anomaly detection strategies in unsupervised settings [3], [6]. Table III summarizes the detection statistics and comparative performance of each method.

TABLE III. PERFORMANCE OF ANOMALY DETECTION

Anomaly Detection Method	**Anomalies Detected**	**Overlap with Other Method**	**Jaccard Index**
Isolation Forest	26	18	0.53
Local Outlier Factor	26	18	0.53
Combined (IF ∩ LOF)	18	18	1.00

As illustrated in Fig. 5, a box plot comparison revealed that anomalous claims exhibited a slightly lower median and tighter interquartile range than the normal group. This contradicts the common assumption that fraudulent claims are always high-cost outliers and instead suggests that irregularities may also stem from atypical billing combinations or low-insurance procedures.

The histogram in Fig. 6 further reinforces this finding. While normal claims displayed a heavy right skew due to a long tail of high-cost cases, anomalous claims were mostly concentrated in the lower-cost spectrum. These results emphasize that the proposed embedding framework detects fraud through complex patterns that extend beyond simple monetary thresholds.

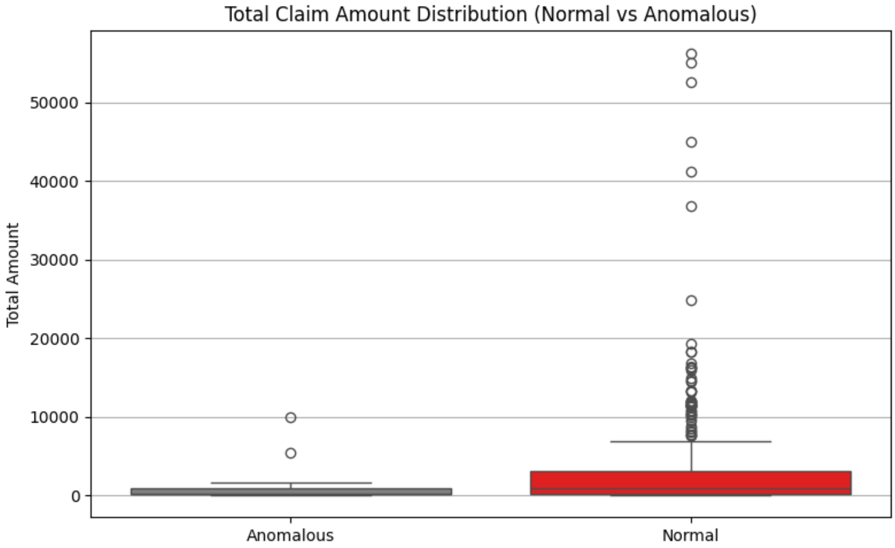

Fig. 5. Box plot of total claim amounts for normal vs. anomalous claims

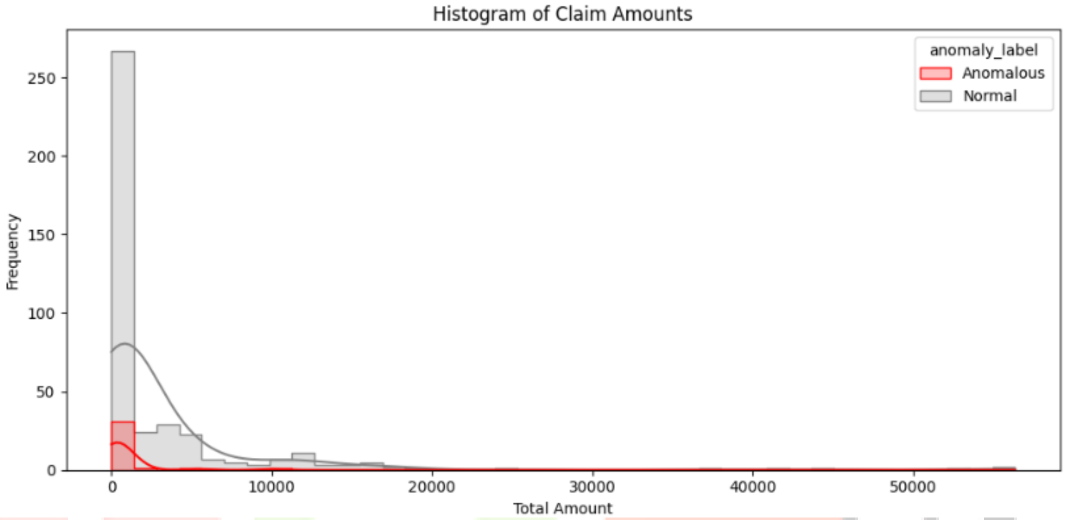

Fig. 6. Histogram of total claim amounts for normal vs. anomalous claims with overlaid density estimates

3) *Feature Correlation Analysis*

To explore which structured features most strongly contributed to anomaly detection, Pearson correlation coefficients were computed between feature values and anomaly scores. Billing amount exhibited the strongest positive correlation, consistent with known fraud risks. Conversely, insurance coverage ratio showed a negative correlation, indicating underinsured claims may be more susceptible to fraud.

Moderate correlations were also observed with the number of diagnoses and submission priority level. These findings validate the inclusion of structured attributes in the multimodal model. The correlation matrix is visualized in Fig. 7.

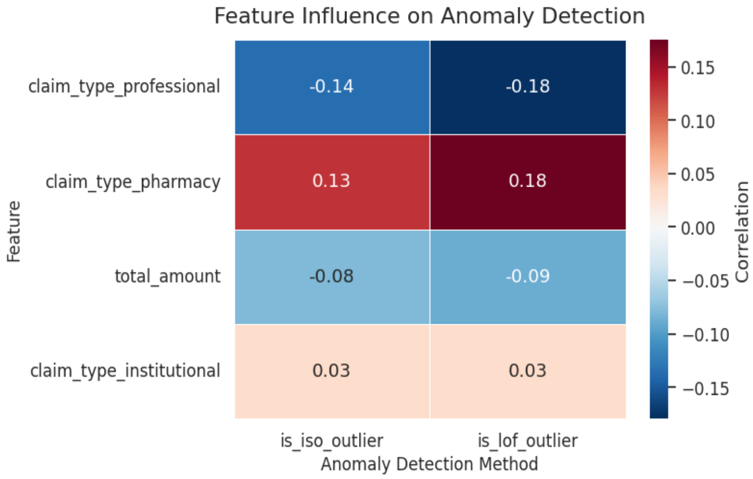

Fig. 7. Feature Correlation Matrix

4) *Provider-Level Fraud Insights*

Beyond claim-level detection, provider-level analysis was performed by aggregating the number of anomalous claims associated with each provider. Only high-confidence anomalies—those jointly flagged by both Isolation Forest and LOF—were considered. As shown in Fig. 8, several providers exhibited disproportionately high counts of anomalous claims. Notably, Provider A was responsible for six such claims, the highest among the set. Providers B, C, and D each reported between four and five flagged submissions. These patterns suggest systematic irregularities potentially linked to billing inflation, duplicate procedures, or low patient diversity—common markers in fraud audits [7]. This provider aggregation serves as a triage mechanism, enabling auditors to prioritize investigations based on anomaly density. Even when individual claims are only marginally suspicious, repeated patterns across a provider's portfolio strengthen the case for further review.

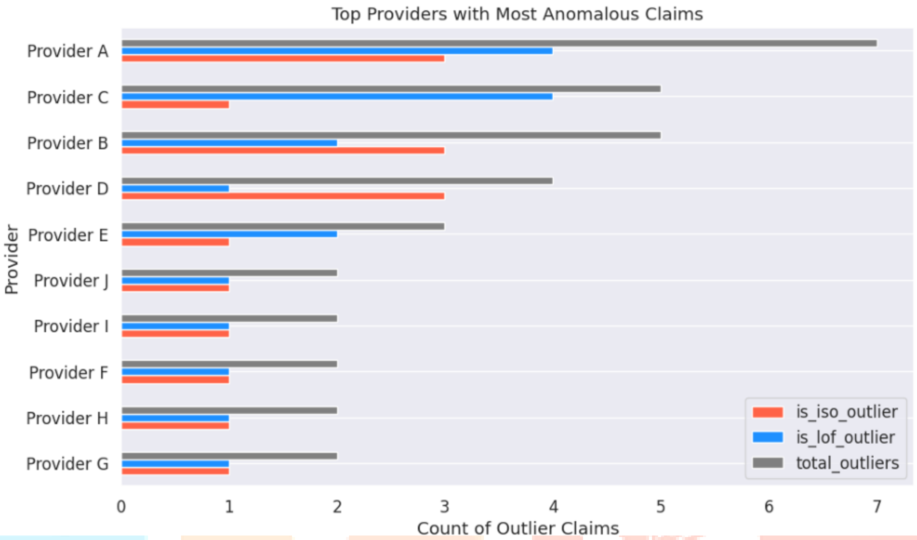

Fig. 8. Top 10 Providers by Anomaly Count

5) *Limitations and Interpretability*

While the proposed unsupervised system performs effectively across multiple dimensions, several limitations must be acknowledged:

- Synthetic Data Dependency: Although Synthea-generated data offers realistic EHR simulation, it may not fully capture the complexity of real-world fraud. Evaluation on actual claims data is necessary to validate model generalizability [16].
- Model Interpretability: Heterogeneous GNNs and deep semantic embeddings, though powerful, remain opaque. Tools such as GNNExplainer or SHAP could enhance transparency by identifying influential features or graph paths [4].
- Lack of Temporal Modeling: The current system operates on static snapshots. However, fraudulent behavior often unfolds temporally. Incorporating temporal GNNs or sequential embeddings may improve detection of time-evolving fraud patterns.

Despite these challenges, the framework demonstrates scalable and label-free fraud analysis and delivers meaningful results at both the micro (claim) and macro (provider) levels.

V. CHALLENGES AND APPLICATIONS.

Real-world deployment of the proposed framework introduces several open challenges. Scaling the heterogeneous graph architecture to handle millions of claims will require optimization strategies such as subgraph sampling, batched training, and distributed inference. The absence of labelled ground truth in healthcare claims complicates evaluation, requiring reliance on anomaly scores and agreement metrics. Additionally, interpretability remains a concern, as both deep embeddings and graph-based inferences may be opaque to clinical stakeholders. Secure and compliant handling of sensitive healthcare data is also essential for adherence to privacy regulations and ethical deployment.

In parallel, the framework offers compelling applications across healthcare fraud detection and related domains. Automated triaging of suspicious claims can significantly improve audit efficiency and reduce manual workload. Provider-level risk scoring supports continuous monitoring of systemic irregularities and enables targeted investigations. Owing to its modular, multimodal design, the system is also adaptable to other sectors such as financial services, cybersecurity, and e-commerce. Its integration into claims adjudication platforms or regulatory systems can enhance real-time decision-making and elevate fraud surveillance capabilities at scale.

VI. CONCLUSION

This paper introduces an unsupervised, multimodal framework for healthcare fraud detection that fuses semantic, structural, and relational information to identify anomalous patterns in insurance claims. By integrating Longformer-based clinical text embeddings, structured claim metadata, and a heterogeneous graph neural network modeling inter-entity relationship, the system captures complex fraud signals beyond surface-level indicators. Experimental results on Synthea-generated synthetic data show that the fused 790-dimensional claim representations, when processed using Isolation Forest and Local Outlier Factor, effectively isolate suspicious claims, and flag providers with consistent irregularities. The model's ability to detect nuanced fraud patterns—such as under-insured procedures or provider-specific anomalies—demonstrates its robustness in the absence of labels. Furthermore, provider-level risk scoring enables organizational surveillance and prioritization of high-risk entities. The architecture is scalable, adaptable, and domain-agnostic, with potential applicability to broader anomaly detection tasks in finance, cybersecurity, and public policy systems. Future work will focus on deployment over real-world datasets, improving explainability through model-agnostic interpretation tools, and incorporating temporal modeling to detect evolving fraud schemes over time.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the support of the Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, for providing the infrastructure and academic environment necessary to carry out this research. The authors also thank the developers of the Synthea™ framework for enabling access to realistic synthetic health records used in this study.

REFERENCES

[1] B. Hong, P. Lu, H. Xu, J. Lu, K. Lin, and F. Yang, “Health insurance fraud detection based on multi-channel heterogeneous graph structure learning,” *Heliyon*, vol. 10, pp. 30045–30065, 2024.

[2] A. A. Amponsah, A. F. Adekoya, and B. A. Weyori, “A novel fraud detection and prevention method for health care claim processing using machine learning and blockchain technology,” *Decis. Anal. J.*, vol. 4, Art. no. 100122, 2022.

[3] A. du Preez, S. Bhattacharya, P. Beling, and E. Bowen, “Fraud detection in healthcare claims using machine learning: A systematic review,” *Artif. Intell. Med.*, vol. 160, Art. no. 103061, 2025.

[4] H. De Meulemeester, F. De Smet, J. van Dorst, E. Derroitte, and B. De Moor, “Explainable unsupervised anomaly detection for healthcare insurance data,” *BMC Med. Inform. Decis. Mak.*, vol. 25, pp. 14–33, 2025.

[5] W. Gamaleldin, O. Attayyib, L. Mohaisen, N. Omer, and R. Ming, “Developing a hybrid model based on Convolutional Neural Network (CNN) and Linear Discriminant Analysis (LDA) for investigating anti-selection risk in insurance,” *J. Radiat. Res. Appl. Sci.*, vol. 18, Art. no. 101368, 2025.

[6] Z. Hamid, F. Khalique, S. Mahmood, A. Daud, A. Bukhari, and B. Alshemaimri, “Healthcare insurance fraud detection using data mining,” *BMC Med. Inform. Decis. Mak.*, vol. 24, pp. 112–135, 2024.

[7] J. T. Hancock, R. A. Bauder, H. Wang, and T. M. Khoshgoftaar, “Explainable machine learning models for Medicare fraud detection,” *J. Big Data*, vol. 10, pp. 154–184, 2023.

[8] S. K. Jena, B. Kumar, B. Mohanty, A. Singhal, and R. C. Barik, “An advanced blockchain-based Hyperledger Fabric solution for tracing fraudulent claims in the healthcare industry,” *Decis. Anal. J.*, vol. 10, Art. no. 100411, 2024.

[9] R. K. L. Kennedy, Z. Salekshahrezaee, F. Villanustre, and T. M. Khoshgoftaar, “Iterative cleaning and learning of big highly imbalanced fraud data using unsupervised learning,” *J. Big Data*, vol. 10, pp. 106–125, 2023.

[10] S. Li, W. Chan, B. Yan, Z. Li, S. Zhu, and Y. Yu, “Self-supervised contrastive representation learning for large-scale trajectories,” *Future Gener. Comput. Syst.*, vol. 148, pp. 357–366, 2023.

[11] J. Lu, K. Lin, R. Chen, M. Lin, X. Chen, and P. Lu, “Health insurance fraud detection by using an attributed heterogeneous information network with a hierarchical attention mechanism,” *BMC Med. Inform. Decis. Mak.*, vol. 23, Art. no. 62, 2023.

[12] E. Nabrawi and A. Alanazi, "Fraud detection in healthcare insurance claims using machine learning," *Risks*, vol. 11, pp. 160–176, 2023.

[13] H. K. Prakosa and N. Rokhman, "Anomaly detection in hospital claims using K-Means and linear regression," *Indones. J. Comput. Cybern. Syst.*, vol. 15, pp. 391–402, 2021.

[14] P. Rochner and F. Rothlauf, "Unsupervised anomaly detection of implausible electronic health records: A real-world evaluation in cancer registries," *BMC Med. Res. Methodol.*, vol. 23, pp. 125–138, 2023.

[15] M. Tabassum, S. Mahmood, A. Bukhari, B. Alshemaimri, A. Daud, and F. Khalique, "Anomaly-based threat detection in smart health using machine learning," *BMC Med. Inform. Decis. Mak.*, vol. 24, pp. 347–362, 2024.

[16] Synthea Project, "Synthetic patient generator." [Online]. Available: https://synthetichealth.github.io/synthea/. Accessed: Mar. 2025.

[17] DGL Team, "Deep Graph Library." [Online]. Available: https://www.dgl.ai. Accessed: Mar. 2025.