IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

Using Graph Attention Networks in Healthcare Provider Fraud Detection

SHAHLA MARDANI[1], and HADI MORADI[1,2],(Senior Member, IEEE)
[1]School of Electrical and Computer Engineering, College of Engineering, University of Tehran, 11155-4563, Tehran, Iran
[2]Intelligent Systems Research Institute, SKKU, Suwon 16419, South Korea

Corresponding author: Hadi Moradi (e-mail: moradih@ ut.ac.ir).

ABSTRACT Healthcare fraud increases healthcare expenses for insurers, premiums for policyholders, and dissatisfaction of legitimate patients and causes severe damage to the health system. Therefore, it is critically important to address healthcare fraud detection. Most fraud detection models consider only claims data for analysis. Since committing healthcare fraud can include more than one party, i.e., healthcare providers, physicians, and patients, it is crucial to consider the relationship among them. In this paper, we propose a healthcare provider fraud detection model that applies the effect of the parties' interdependencies on the claims data. It leverages a graph attention network for embedding the relationships and classifies samples using a feed-forward neural network. Our explicit contribution is using the latent information of the interdependency of claims' parties to detect healthcare provider fraud more accurately. The information, along with other features, can identify complex patterns of fraud. We tested our approach on the healthcare provider fraud detection dataset and reached 0.56 recall compared to best available approaches such as GTN with 0.5 and XGBoost with 0.46 recall.

INDEX TERMS Fraud Detection, Graph Attention Network, Graph Embedding, Healthcare.

I. INTRODUCTION

HEALTH ealth expenses are increasing globally [1] [2]. These expenses were US$ 7.8 trillion in 2017, equivalent to 10% of the US GDP. Statistics show that between 2000 and 2017, healthcare costs increased by 3.9 percent annually, while the economy grew by only 3 percent each year [3]. Due to the size of the industry, the significant transaction amount [4], the complexity of the processes, and the variety of services [5] [6], healthcare is tempting and attractive to fraudsters. In 2018, the National Health Care Anti-Fraud Association (NHCAA) conservatively estimated that the healthcare fraud accounted for approximately 3% of the total healthcare expenditures. Based on the report, some government and law enforcement agencies put the loss as high as 10% of annual healthcare expenditures, which is more than $300 billion [7]. The U.S. Sentencing Commission also reported, "8.0% of theft, property destruction, and fraud offenses involved health care frauds" [8]. Committing fraud leads to increased healthcare expenses for insurers, increased premiums for policyholders, dissatisfaction with legitimate patients, and severe damage to the health system [9]. Thus, detecting healthcare fraud is critically essential.

Different fraud schemes are committed by healthcare providers, physicians, beneficiaries, or insurers, such as Kickback schemes [10], Self-referral [11], Doctor shopping [11], Identity fraud [12], Unbundling [13], Upcoding [14], Fake reimbursements [15], waiving patient co-pays, deductibles, or co-insurance [16], Medications without examination [17]. In some schemes, more than one party is involved [15]. For example, in the self-referral schema, two healthcare providers collude to refer patients to each other. Sometimes, it happens between a physician and a healthcare provider.

Many methods have been suggested for detecting healthcare fraud, including rule-based and machine-learning methods. Due to the dependency of rule-based methods on expert knowledge and the high probability of circumvention by fraudsters, machine learning methods have become the typical approach for fraud detection [18]. Using machine learning in solving the problem of fraud detection has several challenges, such as the existence of inaccurate labels, imbalanced data [19] [20] [21], preprocessing overhead [22], and feature selection. These challenges are prevalent in most fraud detection domains. However, in healthcare, we confront other challenges besides these. In healthcare fraud detection, historical data is essential since fraudulent behavior typically happens over time [5] [23]. Also, the variety of services available for a health issue makes it difficult to detect a pattern of fraud [24] [25] [26]. Another essential issue in fraudulent

behavior is the possible collaboration between parties, such as physicians and healthcare providers [15]. To address this issue, graph analytics is a common and valuable tool because data are often interdependent. On the other hand, interdependencies are more understandable and observable in graphs, even for those that have long-range interdependencies [27]. A graph provides a global view of the entire network [27]. Thus, using the graph of parties and accessing its structure brings us more detailed and rich information than considering a single data point. Thus, fraudsters, who try to cover their relationships, would be caught easier using graph networks. Several studies tried to handle this issue using community detection algorithms [28] or label propagation [29]. Although these methods provided a global view and analysis of the graph of parties' relationships, they calculated or investigated relationship features apart from other features. However, interdependent parties have interdependent features, and they should be considered simultaneously [30]. Also, these works could not determine the different importance of the effect of neighbors, whereas neighbors have different impacts. To consider intrinsic and relationship features simultaneously and give different importances to neighbors, we use Graph Attention Networks (GATs) to embed the graph of parties' relationships and develop an end-to-end neural network classifier for healthcare fraud detection. To the best of our knowledge, there is no work taking advantage of Graph ATtention networks (GAT) in healthcare fraud detection to catch the parties' interdependency, considering their intrinsic features.

Our explicit contribution in this paper is considering and including the parties' intrinsic features and their different importances in their interdependency analysis. Our model uses a deep learning-based graph embedding method to embed the information of parties' relationships and their features simultaneously. We also apply an attention mechanism to determine the importance of each relationship. For a more intuitive understanding of the characteristics of our model, we give an example. Consider two clinical diagnostic labs that present healthcare procedure x frequently. One of them has presented the healthcare procedure x for n patients who have been visited by multiple physicians. Another one has presented the healthcare procedure x to the same number of patients who have been visited by two or three physicians. The second case can be an instance of a self-referral scheme in which healthcare providers refer patients to a specific healthcare provider in exchange for a kickback. It is detectable if we know about the relationships and features of each party (such as their presented healthcare procedures). However, the more relationships parties have, the more complicated it is to detect fraudsters. In this situation, determining different importances for neighbors based on their features is helpful.

The experimental result of applying our method to the dataset [31] shows a higher recall than other proposed models in the domain. Our approach has a better recall, i.e. 0.56, compared to the best results of other approaches, i.e. GTN with 0.5. Other approaches such as XGBoost and Random Forest have 0.46 and 0.39, respectively. We chose the recall measure for evaluation because identifying a non-fraudulent provider as fraudulent (type 1 error) is better than not identifying a fraudulent provider as fraudulent (type 2 error) [32].

We tackle the challenge of historical data and the variety of services by using a semantic embedding of healthcare services for provider features.

The model can also be used in other domains where fraud schemes are formed based on the parties' relationships and interdependencies. However, in this paper, we limit our experiments to healthcare provider fraud.

II. RELATED WORK

Many methods have been presented in the domain of healthcare fraud detection. The most common approaches are rule-based and machine-learning methods [18]. Rule-based models like [33] [34] are made based on the knowledge of experts [35]. Their performance is low because the behavior of fraudsters change over time, and they can circumvent the rules [18] [36]. Also, the model performance is limited to the domain knowledge of experts, and due to skewed class distribution and a small number of fraud records compared to a large number of normal records, detecting fraudulent records is difficult [36] [37].

In contrast, machine learning methods are more flexible [18] than other methods. Machine learning methods are divided into three categories: supervised, unsupervised, and hybrid [38]. In supervised methods, labels are available, and it is possible to train models based on the labeled data. Comparatively, in unsupervised methods, labels are not available. Unsupervised learning has the advantage of anomaly detection, which means it is possible to detect new fraudulent behaviors. Hybrid methods benefit from the advantages of supervised and unsupervised methods together.

The most common supervised methods in the healthcare fraud detection are neural network [39], decision tree [40], segmentation [41], random forest [42], xgboost [43], gradient boosting machine [32], logistic regression [44], Support Vector Machine (SVM) [45], generalized linear model [46], multilayer perceptron [47], attention deep learning [48], graph neural network [49], multi-criteria decision analysis [50], Bayesian multinomial latent variable model [51], and graph analytics [52].

Also, the most common unsupervised methods in the healthcare fraud detection are local outlier factor [53], isolation forest [54], association rules [55], K-Nearest Neighbors (KNN) [56], Hidden Markov Models (HMM) [57], one-class SVM [58], Restricted Boltzmann machine (RBM) [59], regression analysis [60], probabilistic programming [61], community detection [62], outlier detection [63], peak analysis [64], neural network [65], spectral analysis [28], Bayesian co-clustering [66], clustering [67], maximal clique enumeration [68], clustering coefficient [18], page rank [69], graph analytics [70], and statistical analysis [71].

Hybrid methods have commonly used SVM [72], clustering [72], and regression [73].

Most works have focused on model selection to find an appropriate model for detecting fraud. However, determining the suitable model to solve a problem depends on the challenges involved in the context of the problem. In the healthcare fraud detection problem, the following challenges are raised, and researchers have chosen different methods to solve them.

- *Historical behavior:*
  In general, fraudulent behavior is fragmented into different claims over time. It means that for detecting fraudulent behavior, earlier claims by a party, such as a service provider or a patient, should be considered [23]. Kose et al. [5] proposed an interactive machine learning method in which features were extracted based on some storyboards constructed by experts. The features are related to the scenario in each storyboard and aggregated claims' features by parties (e.g., the ratio of the number of prescriptions to the distinct number of insured persons). In other words, the variation of a feature over time becomes a feature by itself. Also, Sun et al. [74] employed historical behaviors to detect camouflage in which fraudsters behave like normal [75] by considering the cluster divergence of each patient's hospital admission graph over time.
- *Interdependency of parties:*
  Committing a healthcare fraud can involve more than one party [15]. For instance, a physician can collaborate with a healthcare provider to issue unnecessary claims for different patients [11]. Thus, such collaboration of parties can be viewed as an interdependency feature between parties, represented as a graph, and then analyzed using graph analysis [76]. Graph analysis has been used in different ways for healthcare fraud detection. The researchers in [18] [68] [69] proposed different similarity measures to build graphs in which nodes represent parties and edges between nodes represent similarities between parties. For instance, the similarity between the prescriptions of different physicians is used as a similarity measure between two physicians. If the similarity between two parties passes a threshold, i.e. a minimum similarity, the nodes are connected in the graph and the similarity value is considered as the weight of the edge. The resulting graph can be used for fraud detection. In this way, extra information for the analysis is provided, but the information on relationship and interdependency between parties is not caught.
  The authors in [70] made a graph of relationships between parties and detected frauds by extracting some features from structural features in the graph (e.g., degree of nodes). The extracted features only caught the information of first-order neighbors for the parties. They were not able to model the interdependencies.
  The methods using community detection algorithms, such as [28], could consider more neighbors than those using structural features. Also, label propagation methods, such as [29], had a global view of the graph and easily propagated the effect of different parties' labels on each other. Although these models considered the parties' interdependencies, they only focused on the relationships. They could not consider both intrinsic features and relationship features simultaneously. To apply a relationship's effect, we should consider the parties' intrinsic features. For example, consider a fraudulent heart clinic with two relationships: a children's hospital and another heart clinic. The interdependency between two heart clinics differs from that between a heart clinic and a children's hospital. Thus, propagating fraudulent labels from the heart clinic to these two providers should be different.
- *Variety of services:*
  Fraudsters apply different fraud scenarios using different treatment processes and make it more challenging to be detected compared to single treatment cases. It can also lead to an increased number of features used for fraud detection that increases the difficulty of accurate fraud detection [77].
  In most research, researchers use these various services as features to model the connections between services and find fraudulent behaviors. The simplest and most common method was one-hot coding, which encoded each categorical value as 0 or 1 [78] and made a numerical feature vector. The method has been used in [79] [80]. The vector provided by one-hot encoding is sparse [2], and it would be costly in terms of memory consumption [43], where unique values of categorical features are high. Large and sparse vectors provided by one-hot encoding would lead to the curse of dimensionality and the model performance degrading [81] [82]. Another disadvantage of one hot encoding is failing to capture relevant relationships between similar values [79].
  Johnson and Khoshgoftaar in [81] addressed the problem by semantic embedding methods inspired by word embedding methods presented in [83] and [84] to detect similarities between values that went undetected when using one-hot vectors.

A list of several studies published in healthcare fraud detection from 2010 to 2022, along with the challenges they have solved, are summarized in TABLE 1.

TABLE 1 shows several studies focused on interdependency that have used graph analytics. The non-linear structure of graphs has made them a suitable tool for modeling complex network structures. Conventional graph analyses, such as path analysis, community detection, and centrality analysis, allow us to extract useful information about relationships. However, these analyses rely on extracting the graph's topological characteristics directly from its structure, typically using adjacency matrices, adjacency lists, or similar representations. This direct approach presents challenges when dealing with large-sized graphs, leading to significant computational costs in terms of both memory and time [87] [88]. Also, distributed

TABLE 1. Papers published in healthcare fraud detection from 2010 to 2022, along with the challenges they have solved.

Paper	Approach	Technique	Year	Historical behavior	Interdependency of the parties	Variety of services
[1]	SP	Convolutional Neural Network (CNN)	2022			*
[81]	SP	Gradient Boosting-Based Tree/ Logistic Regression/ Random Forest	2021			*
[43]	SP	Xgboost/ Catboost	2020			*
[48]	SP	Attention Deep Learning	2020			*
[49]	SP	Graph Neural Network	2020			*
[79]	SP	Neural Networks	2019			*
[85]	SP	Combining Genetic Algorithms/SVM	2019			*
[80]	SP	Gradient Boosting-Based Tree/ Logistic Regression/ Random Forest	2018			*
[86]	SP	Gradient Boosting-Based Tree/ Logistic Regression/ Random Forest	2018			*
[6]	SP	Naïve Bayes	2016			*
[52]	SP	Graph Analytics/ Decision Tree	2016			*
[2]	USP	Isolation Forest/ Unsupervised Random Forest/ Local Outlier Factor/ Autoencoders/ KNN	2018			*
[59]	USP	RBM	2018			*
[74]	USP	Graph-Based Density Peak Clustering	2018	*		*
[29]	USP	Graph Analytics	2017		*	
[69]	USP	Page Rank	2017			*
[5]	USP	Interactive Machine Learning	2015	*		*
[28]	USP	Graph Analytics, Spectral Analysis	2013		*	

(SP: Supervised, USP: Unsupervised).

and parallel processes are not simple and low-cost in this area because nodes are connected through edges, and the nodes distributed in different parts often need each other's information [89]. Graph embedding is an appropriate solution that converts sparse and high-dimensional graphs into a dense and low-dimensional feature vector. The feature vector can be used in machine learning analysis [87].

The common classification of graph embedding methods includes the following three categories [90] [91]:

- Matrix factorization-based: An adjacency matrix can be assumed as the vector representation of a graph. These methods calculate a low dimensional estimation of the adjacency matrix (or another matrix such as the Laplacian matrix), e.g., by singular value decomposition (SVD). Among these methods, we can mention Distributed Large-scale Natural Graph Factorization [92], HOPE [93], and Laplacian Eigenmaps [94].
- Random walk-based: In graph embedding, obtaining the structural information of the graph is important. The adjacency matrix only provides us with the information about the first-order neighbors. Random walk-based methods, inspired by word representation methods such as Word2Vec [95], use random walks to find the neighbors of a node and embed their information into the corresponding feature vector. In this way, the structural information is provided. DeepWalk [96] and Node2Vec [97] are two common methods using this approach.
- Deep learning-based: Unlike random walk-based methods that embed the graph locally, in deep learning-based methods, embedding is done globally. For example, SDNE [98] and DNGR [99] used autoencoder networks to reduce the dimensions of the adjacency matrix. These methods receive neighbors of each node globally (for example, in the form of an adjacency matrix) as input, which could lead to high time costs and high memory consumption for large graphs [100]. Graph convolutional networks (GCN) [101] solved this problem by defining convolutional layers and aggregating neighbor's information (messages) in several iterations. This aggregation was done only on local neighbors and caused the model's scalability [100]. GCN aggregates the neighbors' information by averaging their messages. Another type of graph embedding called GraphSAGE [102] used different approaches for message aggregation (averaging, pooling, and LSTM). Also, graph attention networks (GAT) [103] considered different importance for the influence of each of the neighbors by using the attention mechanism [104] in deep networks. The main idea was that the information provided by all nodes are not equally important. In this way, it considered different weights for the information of neighbor nodes.

Matrix factorization-based methods provide a good estimation due to embedding graph nodes globally but have a high computational cost in terms of time and memory [87]. In contrast, random walk-based methods that embed graph nodes locally and deep learning-based methods that learn embeddings by aggregating information from local neighborhoods can be more scalable than matrix factorization-based methods. The fundamental nature of matrix factorization involves global computations that are inherently computationally expensive. Even with efficient representations like adjacency lists, the computational steps required for factorization remain more intensive compared to the localized, scalable methods used in deep learning and random walk-based methods.

Although both deep learning and random walk-based methods are scalable and usually do not require feature engineering [90], deep learning-based methods can catch complex relationships and high-order structures compared with random walk-based methods because of their layer-wise global aggregation, while the performance of random walk-based methods depends on local nodes and the number of random walks [90].

In this paper, we propose a method based on the GAT to catch the interdependency between parties. Using GAT, we can map relationships in a feature vector and reduce the complexity of analyzing relationships. Also, it can consider

TABLE 2. The number of records in the fraud dataset.

Dataset	Number of Records
Beneficiary	138564
Inpatient	40514
Outpatient	517737
Provider Label	5410

TABLE 3. Healthcare Provider Fraud Detection Analysis dataset published in Kaggle.

Beneficiary Dataset	Inpatient Dataset	Outpatient Dataset	Label Dataset
Beneficiary Id	Beneficiary Id	Beneficiary Id	Provider
Date Of Birth	Claim Id	Claim Id	Potential Fraud
Date Of Death	Claim Start Date	Claim Start Date	
Gender	Claim End Date	Claim End Date	
Race	Provider	Provider	
Renal Disease Indicator	Insurance Claim Amount Reimbursed	Insurance Claim Amount Reimbursed	
State	Attending Physician	Attending Physician	
County	Operating Physician	Operating Physician	
No Of Months-Part A Coverage	Other Physician	Other Physician	
No Of Months-Part B Coverage	Claim Diagnosis Code 1	Claim Diagnosis Code 1	
Chronic Condition - Alzheimer	Claim Diagnosis Code 2	Claim Diagnosis Code 2	
Chronic Condition - Heart Failure	Claim Diagnosis Code 3	Claim Diagnosis Code 3	
Chronic Condition - Kidney Disease	Claim Diagnosis Code 4	Claim Diagnosis Code 4	
Chronic Condition - Cancer	Claim Diagnosis Code 5	Claim Diagnosis Code 5	
Chronic Condition - Obstructive Pulmonary	Claim Diagnosis Code 6	Claim Diagnosis Code 6	
Chronic Condition - Depression	Claim Diagnosis Code 7	Claim Diagnosis Code 7	
Chronic Condition - Diabetes	Claim Diagnosis Code 8	Claim Diagnosis Code 8	
Chronic Condition - Is Chemic Heart	Claim Diagnosis Code 9	Claim Diagnosis Code 9	
Chronic Condition - Osteoporosis	Claim Diagnosis Code 10	Claim Diagnosis Code 10	
Chronic Condition – Rheumatoid Arthritis	Claim Procedure Code 1	Claim Procedure Code 1	
Chronic Condition - Stroke	Claim Procedure Code 2	Claim Procedure Code 2	
Inpatient Annual Reimbursement Amount	Claim Procedure Code 3	Claim Procedure Code 3	
Inpatient Annual Deductible Amount	Claim Procedure Code 4	Claim Procedure Code 4	
Outpatient Annual Reimbursement Amount	Claim Procedure Code 5	Claim Procedure Code 5	
Outpatient Annual Deductible Amount	Claim Procedure Code 6	Claim Procedure Code 6	
	Admission Date	Deductible Amount Paid	
	Claim Admit Diagnosis Code	Claim Admit Diagnosis Code	
	Deductible Amount Paid		
	Discharge Date		
	Diagnosis Group Code		

both intrinsic features and relationships simultaneously.

While GAT has been utilized for various purposes such as fake news detection [105], Play to Earn Massively Multiplayer Online Role-Playing Games (P2E MMORPG) chargeback fraud detection [106], account takeover [107], and transaction fraud detection [108], its application for healthcare fraud detection and identifying interdependencies between parties is novel and unprecedented.

III. PROPOSED METHOD

A. DATA

In this study, we used the Healthcare Provider Fraud Detection Analysis dataset published in Kaggle [31]. The dataset consists of four sets: beneficiary, inpatient claims, outpatient claims, and provider fraud labels. TABLE 2 presents the number of records in the dataset, and TABLE 3 shows the features available in the dataset. As shown in TABLE 3, there is a set of inpatient and outpatient claims. It includes a list of different healthcare providers and their services to patients by physicians (attending, operating, and other physicians). Each claim contains a set of diagnosis codes and a set of healthcare service codes registered according to the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) [109] and shown in the Claim Diagnosis Codes and the Claim Procedure Codes. In the dataset, the providers have been labeled fraudulent or not, while the claims do not have such a label.

B. FEATURE ENGINEERING

In our problem, there are three participants: healthcare providers, physicians (attending, operating, or other), and beneficiaries (patients). The dataset only has explicit features (e.g., age) for beneficiaries. Other participants do not include such features. However, we can find lots of information from their claims. Due to the fragmented behavior in healthcare fraud, it is essential to pay special attention to the history of claims for each participant. Since, in this study, we focused on the fraud committed by providers, we limited our feature selection to providers' features. Also, for simplicity, we limited our features to claim procedure codes and left the use of other features for future work. Claim procedure codes can present providers' features such as their specialties implicitly.For using claim procedure codes as features, we used the HcpcsVec embedding technique [81] in three steps:

Inpatient Claims

Claim ID	PROVIDER	PRC1	PRC2
CLM37694	PRV51244	66	2724
CLM53301	PRV51244	66	4019
CLM80237	PRV51244	66	4019
CLM59012	PRV51244	66	4019
CLM40186	PRV51244	66	4019
CLM44431	PRV51244	66	4019
CLM78910	PRV57335	66	4019
CLM69968	PRV55039	66	4139
CLM49249	PRV51244	3241	496
CLM78037	PRV51244	3722	66
CLM56910	PRV51244	3722	
CLM71039	PRV51244	3722	
CLM33249	PRV51244	3772	4019
CLM44667	PRV51244	3772	
CLM49820	PRV55039	3950	4019
CLM39780	PRV55039	3950	
CLM45207	PRV57335	6859	496

Outpatient Claims

Claim ID	PROVIDER	PRC1	PRC2	PRC3	PRC4
CLM483763	PRV51244	66	496		
CLM418024	PRV57335	66	7820	412	7840
CLM155068	PRV55039	604	4019		
CLM524797	PRV57335	4516			
CLM720107	PRV55039	9952			

Combined Dataset

Claim ID	PROVIDER	PRC1	PRC2	PRC3	PRC4
CLM37694	PRV51244	66	2724		
CLM53301	PRV51244	66	4019		
CLM80237	PRV51244	66	4019		
CLM59012	PRV51244	66	4019		
CLM40186	PRV51244	66	4019		
CLM44431	PRV51244	66	4019		
CLM78910	PRV57335	66	4019		
CLM69968	PRV55039	66	4139		
CLM49249	PRV51244	3241	496		
CLM78037	PRV51244	3722	66		
CLM56910	PRV51244	3722			
CLM71039	PRV51244	3722			
CLM33249	PRV51244	3772	4019		
CLM44667	PRV51244	3772			
CLM49820	PRV55039	3950	4019		
CLM39780	PRV55039	3950			
CLM45207	PRV57335	6859	496		
CLM483763	PRV51244	66	496		
CLM418024	PRV57335	66	7820	412	7840
CLM155068	PRV55039	604	4019		
CLM524797	PRV57335	4516			
CLM720107	PRV55039	9952			

Provider-Service Occurrences

PROVIDER ID	66	412	496	604	2724	3241	3722	3772	3950	4019	4139	4516	6859	7820	7840	9952
PRV51244	8	0	2	0	1	1	3	2	0	6	0	0	0	0	0	0
PRV57335	2	1	1	0	0	0	0	0	0	1	0	1	1	1	1	0
PRV55039	1	0	0	1	0	0	0	0	2	1	2	0	0	0	0	1

Normalized Provider-Service Occurrences

PROVIDER ID	66	412	496	604	2724	3241	3722	3772	3950	4019	4139	4516	6859	7820	7840	9952
PRV51244	1	0	1	0	1	1	1	1	0	1	0	0	0	0	0	0
PRV57335	0.14	1	0.5	0	0	0	0	0	0	0	0	1	1	1	1	0
PRV55039	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	1

FIGURE 1. Embedding of procedure codes as provider features. Each claim can include multiple procedure codes listed as PRC1 to PRC4. The combined dataset is created by combining Outpatient and inpatient datasets. Then, the number of occurrences of each claim procedure for each healthcare provider is calculated (Provider-Service Occurrences dataset). The final dataset is obtained by normalizing the Provider-Service Occurrences dataset.

1) Combined Dataset: Combine inpatient and outpatient datasets.
2) Provider-Service Occurrences: Group claims by each provider and sum up the number of claims for each claim procedure.
3) Normalized Provider-Service Occurrences: Use Min-Max Scaling [110] for normalizing Provider-Service Occurrences.

In other words, we consider the number of occurrences of each claim procedure for a healthcare provider as its features. FIGURE 1 shows the embedding of procedure codes (PRCs) as provider features. PRCs are claim procedure code features whose corresponding values show ICD-9-CM procedure codes. Each claim can include multiple PRCs that are listed as PRC1 to PRC4. After combining outpatient and inpatient datasets, we group all claims by each provider and sum up the number of claims for each claim procedure (provider-Service Occurrences). Finally, we normalize the resulting dataset by Min-Max scaling. Since there are 1324 unique procedure codes in the dataset, in addition to a null value, the number of features equals 1325.

C. GRAPH STRUCTURE

In healthcare, fraud is committed by one participant alone or by more than one. In these situations, there are two types of relationships between participants that form fraud scenarios:

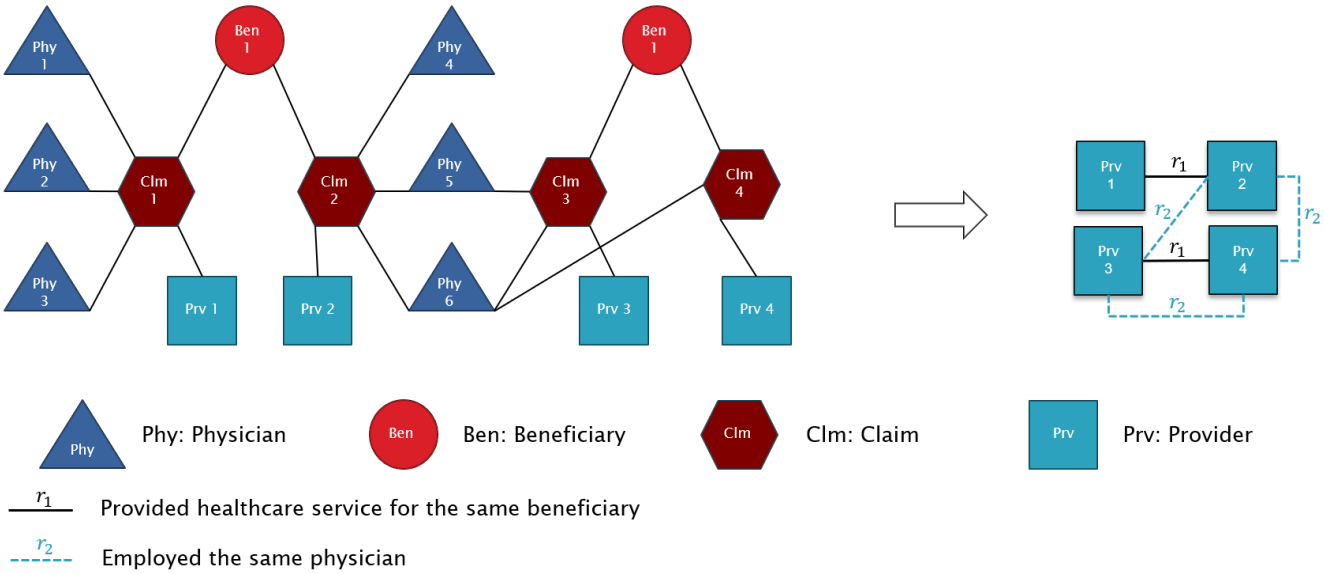


FIGURE 2. **A sample of the relationship between different participants and its transformation into a network of providers. Left) A heterogeneous graph of physicians, providers, and beneficiaries. Right) the desired target graph.**

1) When a fraudster commits fraud alone, he/she will repeat it in different situations with different participants.
2) When fraudsters work together, they trust each other and try to cover each other to conceal their fraud.

For example, consider a fraudster who misuses the ID card of a beneficiary and receives some services from a healthcare provider. He/she may repeat it with different healthcare providers to prevent being detected. Another example can be a healthcare provider and a beneficiary colluding with each other and billing unrealistic claims for an insurance company. These relationships may exist between every two or three participants: Provider and beneficiary, beneficiary and physician, physician and physician, or between beneficiary, physician, and provider. Therefore, if a participant has committed fraud, other participants with a relationship with it would be suspicious, which is a case of interdependency. Since all claims are billed by a provider to an insurance company, whether the provider is aware of the fraud of other participants or not, it is responsible for the fraud. Thus, for capturing the interdependency, we consider two types of relationships between providers: (i) providers who have provided services to the same patient and (ii) providers who have employed the same physician. We model their relationships by an unweighted graph *G=(V, R)* where *V* represents providers, and *R* represents relations between them. Relations (edges) are presented by (v_i, r_k, v_j) where r_k is the relation type (edge type): (i) provided services to the same patient, (ii) employed the same physician.

We only keep the history up to thirty days. In other words, we create the edges if the difference of the start date in the claims that have the same participant (physician/ beneficiary) is at most 30 days. FIGURE 2 shows a sample of the relationship between different participants and its transformation into a network of providers.

On the left side of FIGURE 2, a heterogeneous graph is shown. The graph shows the relationship between parties (physicians, providers, and beneficiaries). Some parties are connected to more than one claim. For example, beneficiary 1 (Ben1) is connected to two claims (Clm1 and Clm2), and each of the claims is connected to different healthcare providers (Prv1 and Prv2). It means the beneficiary has received healthcare services from the providers (Prv1 and Prv2). Thus, in our target graph (right side of FIGURE 2), two nodes, Prv1 and Prv2, are connected because of the same beneficiary (edge type r1). It also works for the same physicians. Prv2 and Prv3 are connected (edge type r2) in the target graph (right side) due to two shared physicians (Phy5 and Phy6) in the primitive graph (left side).

D. GRAPH EMBEDDING

In this section, we explain our supervised model based on the Graph ATtention Network (GAT) [103] for healthcare provider fraud detection. In our graph, there are different edge types. At first, for each type of edge, we design the transformation matrix M_k to project the features of nodes into the feature space based on the edge type k. The projection process can be shown as follows:

$$h_i' = M_k . h_i \tag{1}$$

in which h_i and h_i' are the original and projected features of node i, respectively. To learn the weight of different nodes, we perform self-attention mechanism [104]. Given a node pair *(i, j)* connected via edge type k, the attention ${e_{ij}}^k$ calculates the importance of node j's features to node i. It can be formulated as follows:

$${e_{ij}}^k = att(h_i', h_j', k) \tag{2}$$

Here, *att* is a deep neural network which performs the attention. We use masked attention [104] in the way that ${e_{ij}}^k$ is only computed for thenodes $j \in N_i^k$, where N_i^k is the set of neighbors of node i in the graph based on the edge type k.

We consider the first-order neighbors of i (including i) in all our experiments. To make comparable coefficients across different nodes, we normalize them using the softmax function:

$$ {\alpha_{ij}}^k = softmax_j({e_{ij}}^k) = \frac{exp(\sigma(a_k^T.[h_i' \| h_j']))}{\sum_{l \in N_i^k} exp(\sigma(a_k^T.[h_i' \| h_l']))} \quad (3) $$

in which $._T$ represents transposition, $\|$ denotes the concatenate operation, σ denotes the activation function, and a_k is the attention vector for edge type k. As we can see from equation (3), the *(i,j)*'s features influence on the weight coefficient. Because of the concatenation order in the numerator and the difference between their neighbors that makes the normalized term (denominator), the weight coefficient ${\alpha_{ij}}^k$ is asymmetric and makes different contributions to each node. Then, we aggregate the neighbor's projected features of node i by the corresponding coefficients to calculate the edge type-based embedding as follows:

$$ {z_i}^{k'} = \sigma(\sum_{j \in N_i^k} {\alpha_{ij}}^k.h_j') \quad (4) $$

in which z_i^k is the learned embedding of node i for the edge type k. Multi-head attention [104] is used to stabilize the learning process of self-attention. We execute M independent attention mechanisms (equation (4)) and concatenate their features as follows:

$$ {z_i}^{k'} = \Big\|_{m=1}^{M} \sigma(\sum_{j \in N_i^k} {\alpha_{ij}}^k.h_j') \quad (5) $$

In our experiment, we applied two GAT layers. The first layer aggregates the messages received from their immediate neighbors and sends a graph representation to the next layer. The second layer updates the current representation of the graph by aggregating the messages received from their neighbors. As a result, each layer increases the receptive field by one hop:

$$ {z_i}^{k''} = \sigma(\sum_{j \in N_i^k} {\alpha_{ij}}^{k^{(l_1)}}.h_j'^{(l_1)}) \quad (6) $$

in which ${a_{ij}}^{k^{(l_1)}}$ and $h_j'^{(l_1)}$ denote the weight coefficient and the projected feature of node j in the first GAT layer, respectively.

After GAT layers, we use a feed-forward layer:

$$ z_i = \sigma(\sum_{i \in V} {z_i}^{k''}.{w_i}^k + {b_i}^k) \quad (7) $$

in which ${w_i}^k$ denotes weight and ${b_i}^k$ is bias in edge-type k embedding. Given the K edge types, we will have K groups of node embeddings, denoted as z^k , that are concatenated as follows:

$$ {z_i}^{k'} = \Big\|_{k=1}^{K} {z_i}^k \quad (8) $$

TABLE 4. The model hyperparameters.

Hyperparameter	Tested value(s)
the dimension of the attention vector	[20, 28]
The number of attention heads	[1, 4, 8]
Learning rate	[0.001, 0.01]
Regularization	[0.0005, 0.001, 0.005]
the number of epochs	100
Dropout	0.6

Finally, the resulting feature space is passed through a feed-forward neural network and prediction y_i' is calculated:

$$ y_i' = \sigma(\sum_{i \in V} {z_i}^{k''}.{w_i}^k + {b_i}^k) \quad (9) $$

FIGURE 3 presents the overall framework of the proposed model. According to the previous section, we make a graph of healthcare providers' relationships, including two edge types: the same beneficiary (*r1*) and the same physician (*r2*). As seen in FIGURE 3, we have a two-layer GAT embedding and a feed-forward layer for each edge type. Resulting embeddings from two edge types are concatenated and used as an input for a feed-forward layer. We Minimize the Cross-Entropy equation represented in (10) between the ground truth and the prediction:

$$ CE = \frac{1}{V} \sum_{i \in V} y_i \log(y_i') + (1 - y_i) \log(1 - y_i') \quad (10) $$

where y_i denotes the actual label for node i and y_i' denotes the predicted label for node i.

IV. EXPERIMENTAL RESULTS

A. IMPLEMENTATION DETAIL

We initialized the parameters randomly and optimized the model using Adam [111]. Hyperparameters and their values are shown in TABLE 4 . For some hyperparameters, a range of values was considered and tested. We split the dataset into a training set, a validation set, and a test set with a ratio of 60%, 20%, and 20%, respectively, to ensure fairness.

Our experiments were divided into two stages. At first, we performed our model with different hyperparameters to find the most appropriate ones. In this stage, we also implemented a Graph transformer network (GTN) [112], a deep learning-based graph embedding method for heterogeneous graphs to evaluate the performance of GATs in our model. In the second stage, we compared the effectiveness of our model with some conventional algorithms including, Random Forst [113], XGBoost [114], and Logistic Regression [115], to check if the complexity in our model create an advantage over these algorithms. We chose these algorithms because Akbar et al. [32] and Herland et al. [19] have shown that these algorithms have the best performance on imbalanced datasets.

It should be noted that we tested our approach on publicly available datasets in which the inter-parties relations are available. Thus, we could not test our approach on datasets such as Medicare Part B [116]. Therefore, we compared our model with the algorithms on [32] (Random Forst and XGBoost),

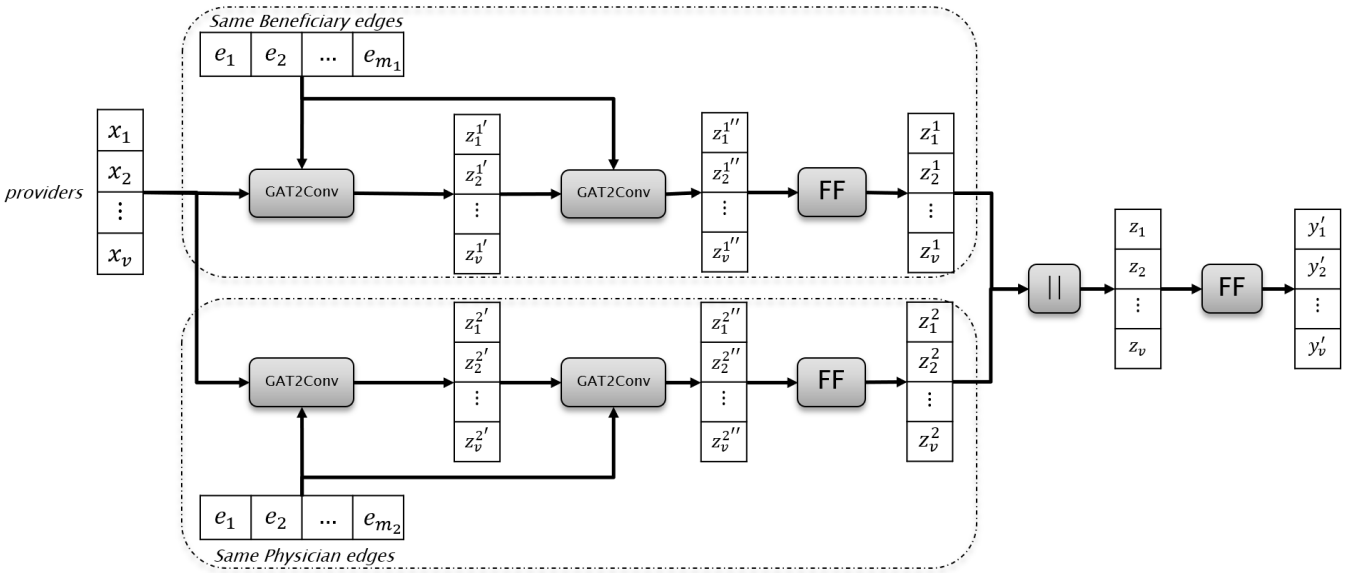


FIGURE 3. The overall framework of the proposed model.A two-layer GAT embedding and a feed-forward layer are considered for each edge type. The resulting embeddings are concatenated and used as an input for a feed-forward layer.

which has the best performance on the database. For more certainty, we added the evaluation of the algorithm logistic regression, which is shown as the best performance algorithm on the Medicare Part B in [19].

B. RESULT ANALYSIS

For evaluation, we calculated the measure of precision, recall, F1-score, and AUC. Since, the performance of our model and GTN are depended on their hyperparameters, we tested these models with different hyperparameters to find the best performance for each one. In TABLE 5 we have summarized the best results of our model and GTN model on the healthcare provider fraud detection dataset. Also, in FIGURE 4 to 7 we have presented precision, recall, F1-score, and AUC respectively. Considering different hyperparameters and measures, our GAT-based model performs better than the GTN. As seen the GAT-based model (attention dimension=20, head=1, LR=0.001, Regularization= 0.005), the GAT-based model (attention dimension=28, head=4, LR=0.01, Regularization = 0.0005), the GAT-based model (attention dimension=28, head=8, LR=0.001, Regularization = 0.001), and the GAT-based model (attention dimension=28, head=4, LR=0.01, Regularization = 0.0005), show the best precision, recall, F1-score and AUC respectively.

Although GTN can recognize more complex relationships and dependencies due to the use of additional transformer-like layers, it has not performed well here. By investigating the precision measure on the train set, we found that the precision of the GTN model was close to 0.98 in all configurations for the train data set, while the precision was not more than 0.5 for validation and test data sets. It can indicate that the model is overfitting because of its inappropriate architecture for this specific task. The result shows that the GAT-based model, with its more straightforward architecture, has generalized and performed better.

The results presented in TABLE 6 show that these models also outperform Random Forest, XGBoost, and Logistic Regression models. However, in our domain, detecting a non-fraudulent provider as fraudulent (type 1 error) is better than not detecting a fraudulent provider as fraudulent (type 2 error) [32]. Therefore, we chose the GAT-based model (attention dimension=28, head=4, LR=0.01, Reg= 0.0005) as our best model because it has the highest recall.

V. CONCLUSION

Healthcare fraud leads to increased healthcare expenses for insurers, increased premiums for policyholders, dissatisfaction of legitimate patients, and severe damage to the healthcare system. Thus, healthcare fraud detection is critically important. In this paper, we have proposed a healthcare provider fraud detection model that uses a GAT to capture the relation and interdependency between participants in claims to find fraudulent behavior. We use both intrinsic and relationship features simultaneously in our model to increase the fraud detection model performance. Our model outperforms other fraud detection methods in this domain with higher recall.

For future work, we consider investigating the effect of using other features (e.g., diagnosis code) and applying feature selection methods. Also, we will use an under/oversampling method due to the fact that fraud detection datasets are normally biased toward majority classes (normal class) and lead to poor performance on the minority class (fraud class). Thus, by under/oversampling, the model can learn to better distinguish between classes and improve overall performance.

Since the model is applicable in every domain where fraud schemes are formed based on the parties' relationships, per-

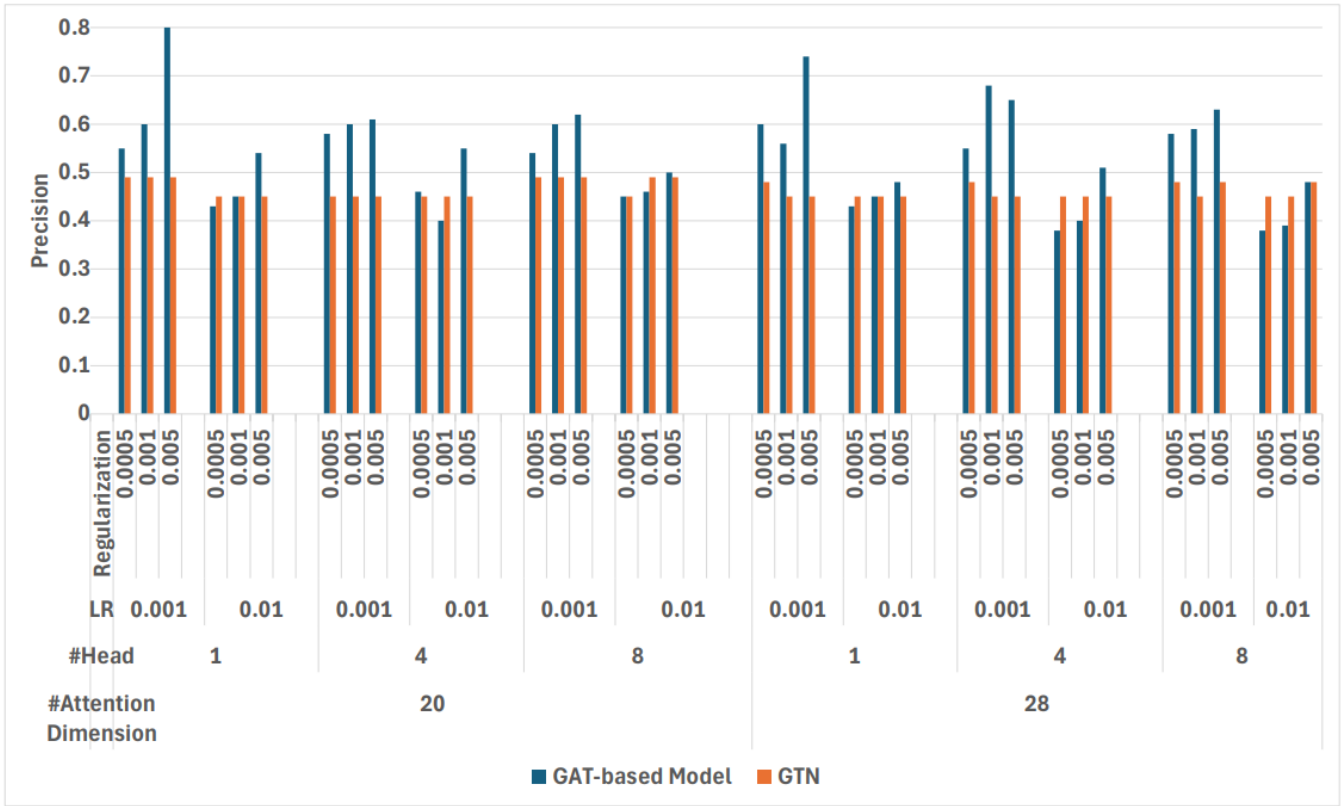


FIGURE 4. **Precision of our model and GTN model on the dataset for different hyperparameters.**

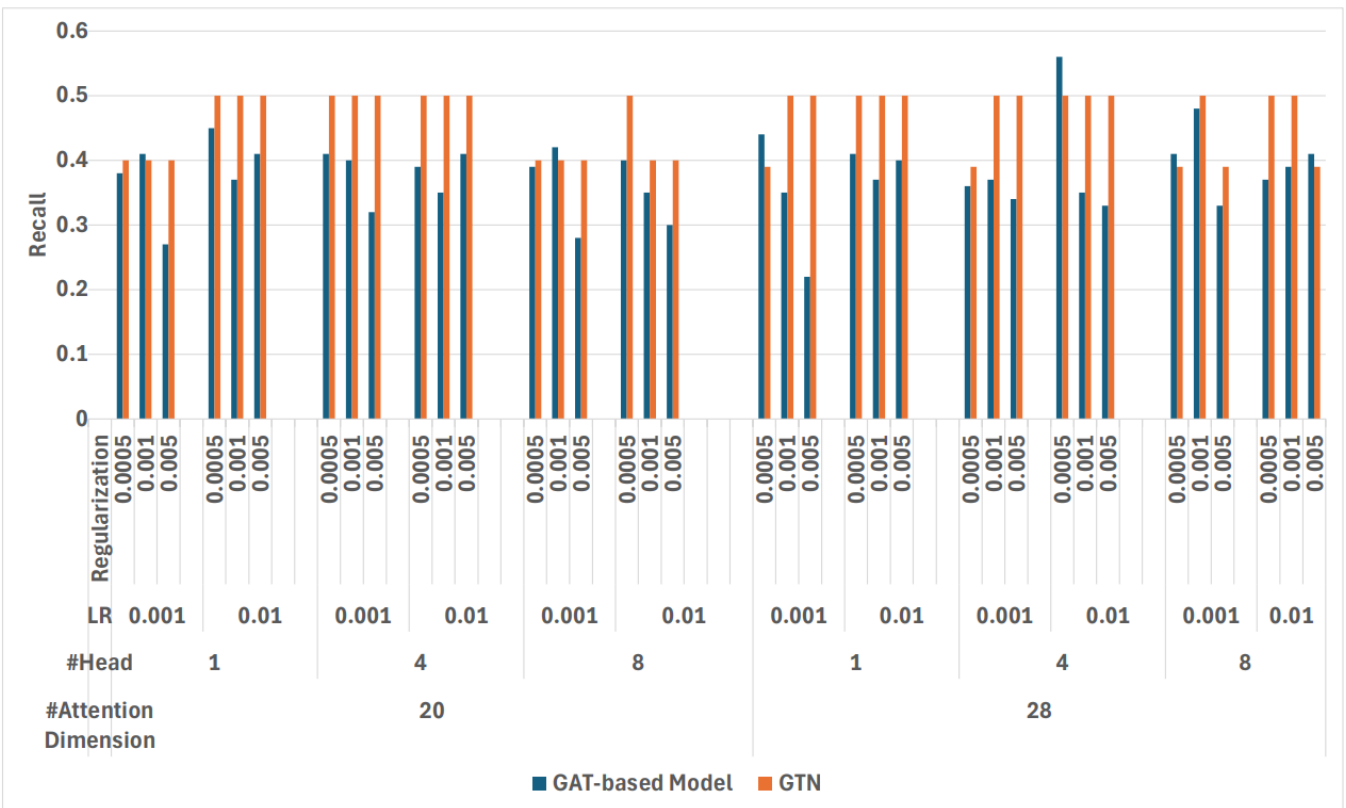


FIGURE 5. **Recall of our model and GTN model on the dataset for different hyperparameters.**

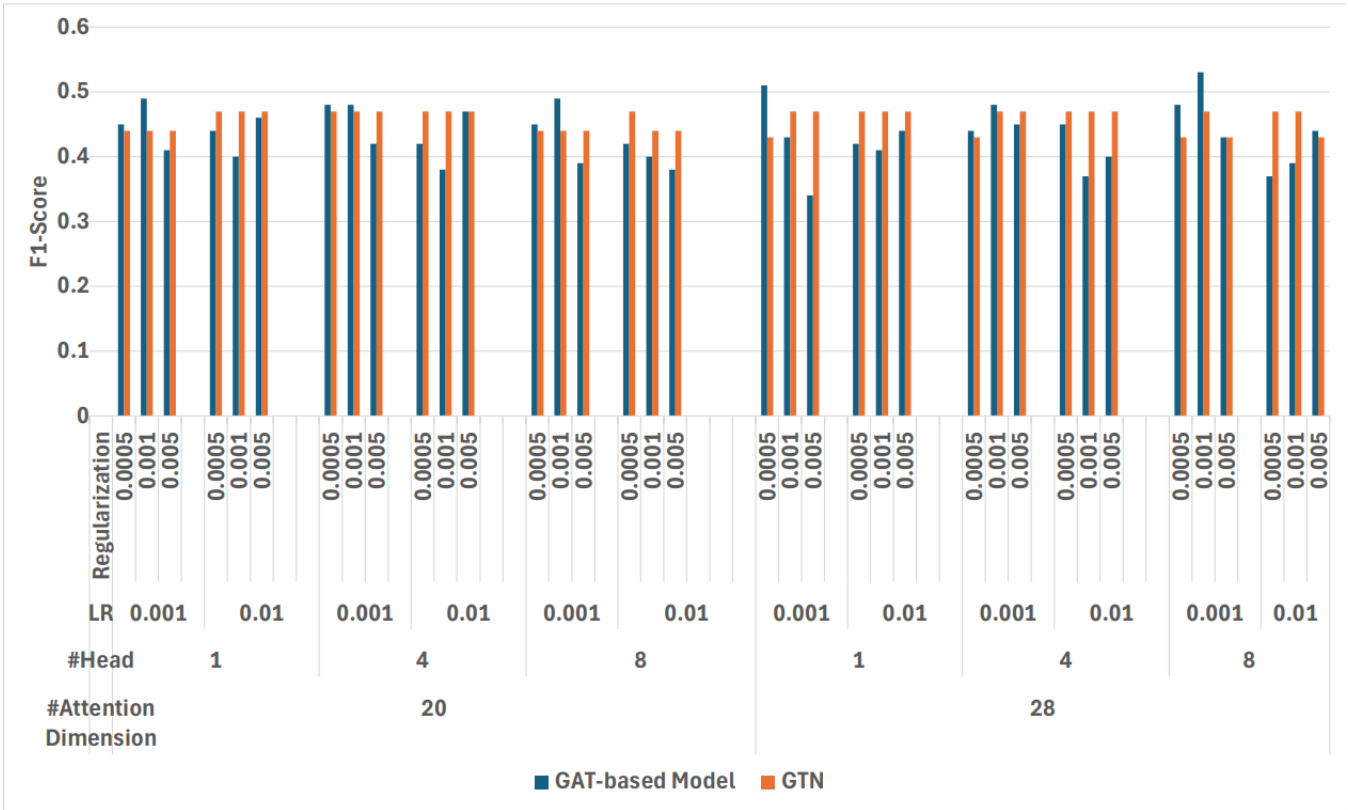


FIGURE 6. F1-score of our model and GTN model on the dataset for different hyperparameters.

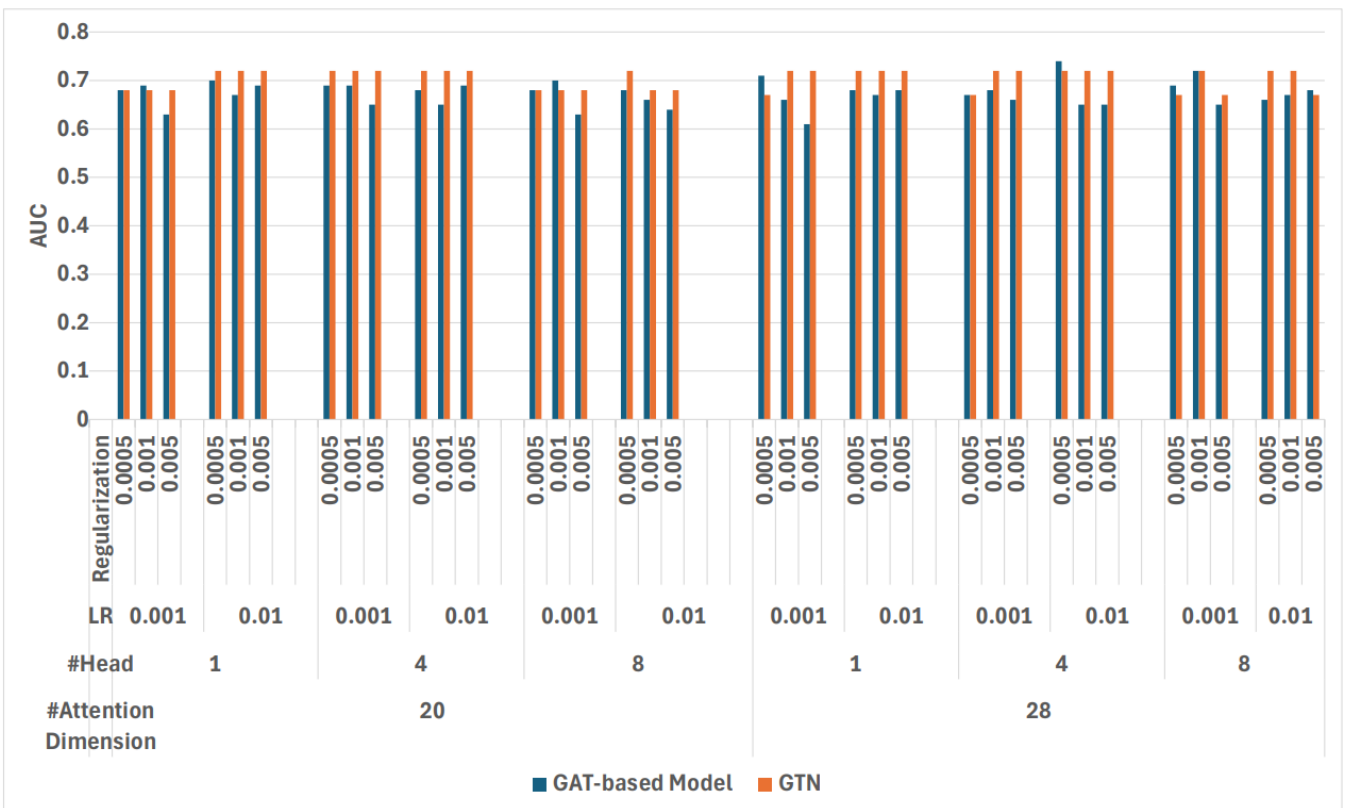


FIGURE 7. AUC of our model and GTN model on the dataset for different hyperparameters.

TABLE 5. The results of performing our model and GTN model on the dataset.

Hyperparameters				GAT-based model				GTN			
#Attention Dimension	#Head	Learning Rate (LR)	Regularization	Precision	Recall	F1-Score	AUC	Precision	Recall	F1-Score	AUC
20	1	0.001	0.0005	0.55	0.38	0.45	0.68	0.49	0.4	0.44	0.68
20	1	0.001	0.001	0.6	0.41	0.49	0.69	0.49	0.4	0.44	0.68
20	1	0.001	0.005	**0.8**	0.27	0.41	0.63	0.49	0.4	0.44	0.68
20	1	0.01	0.0005	0.43	0.45	0.44	0.7	0.45	0.5	0.47	0.72
20	1	0.01	0.001	0.45	0.37	0.4	0.67	0.45	0.5	0.47	0.72
20	1	0.01	0.005	0.54	0.41	0.46	0.69	0.45	0.5	0.47	0.72
20	4	0.001	0.0005	0.58	0.41	0.48	0.69	0.45	0.5	0.47	0.72
20	4	0.001	0.001	0.6	0.4	0.48	0.69	0.45	0.5	0.47	0.72
20	4	0.001	0.005	0.61	0.32	0.42	0.65	0.45	0.5	0.47	0.72
20	4	0.01	0.0005	0.46	0.39	0.42	0.68	0.45	0.5	0.47	0.72
20	4	0.01	0.001	0.4	0.35	0.38	0.65	0.45	0.5	0.47	0.72
20	4	0.01	0.005	0.55	0.41	0.47	0.69	0.45	0.5	0.47	0.72
20	8	0.001	0.0005	0.54	0.39	0.45	0.68	0.49	0.4	0.44	0.68
20	8	0.001	0.001	0.6	0.42	0.49	0.7	0.49	0.4	0.44	0.68
20	8	0.001	0.005	0.62	0.28	0.39	0.63	0.49	0.4	0.44	0.68
20	8	0.01	0.0005	0.45	0.4	0.42	0.68	0.45	0.5	0.47	0.72
20	8	0.01	0.001	0.46	0.35	0.4	0.66	0.49	0.4	0.44	0.68
20	8	0.01	0.005	0.5	0.3	0.38	0.64	0.49	0.4	0.44	0.68
28	1	0.001	0.0005	0.6	0.44	0.51	0.71	0.48	0.39	0.43	0.67
28	1	0.001	0.001	0.56	0.35	0.43	0.66	0.45	0.5	0.47	0.72
28	1	0.001	0.005	0.74	0.22	0.34	0.61	0.45	0.5	0.47	0.72
28	1	0.01	0.0005	0.43	0.41	0.42	0.68	0.45	0.5	0.47	0.72
28	1	0.01	0.001	0.45	0.37	0.41	0.67	0.45	0.5	0.47	0.72
28	1	0.01	0.005	0.48	0.4	0.44	0.68	0.45	0.5	0.47	0.72
28	4	0.001	0.0005	0.55	0.36	0.44	0.67	0.48	0.39	0.43	0.67
28	4	0.001	0.001	0.68	0.37	0.48	0.68	0.45	0.5	0.47	0.72
28	4	0.001	0.005	0.65	0.34	0.45	0.66	0.45	0.5	0.47	0.72
28	4	0.01	0.0005	0.38	**0.56**	0.45	**0.74**	0.45	0.5	0.47	0.72
28	4	0.01	0.001	0.4	0.35	0.37	0.65	0.45	0.5	0.47	0.72
28	4	0.01	0.005	0.51	0.33	0.4	0.65	0.45	0.5	0.47	0.72
28	8	0.001	0.0005	0.58	0.41	0.48	0.69	0.48	0.39	0.43	0.67
28	8	0.001	0.001	0.59	0.48	**0.53**	0.72	0.45	0.5	0.47	0.72
28	8	0.001	0.005	0.63	0.33	0.43	0.65	0.48	0.39	0.43	0.67
28	8	0.01	0.0005	0.38	0.37	0.37	0.66	0.45	0.5	0.47	0.72
28	8	0.01	0.001	0.39	0.39	0.39	0.67	0.45	0.5	0.47	0.72
28	8	0.01	0.005	0.48	0.41	0.44	0.68	0.48	0.39	0.43	0.67

TABLE 6. The results of performing our model, Random Forest, XGBoost, and Logistic Regression on the dataset.

Model	Precision	Recall	F1-Score	AUC
GAT-based model (attention dimension=20, head=1, LR=0.001, Reg= 0.005)	**0.8**	0.27	0.41	0.63
GAT-based model (attention dimension=28, head=4, LR=0.01, Reg= 0.0005)	0.38	**0.56**	0.45	**0.74**
GAT-based model (attention dimension=28, head=8, LR=0.001, Reg= 0.001)	0.59	0.48	**0.53**	0.72
Random Forest	0.78	0.39	0.52	0.66
XGBoost	0.67	0.46	0.46	0.68
Logistic Regression	0.7	0.31	0.43	0.62

forming the model on other domains, such as auction frauds and money laundering, will also be considered.

REFERENCES

[1] T. Matschak, C. Prinz, F. Rampold and S. Trang, "Show Me Your Claims and I'll Tell You Your Offenses: Machine Learning-Based Decision Support for Fraud Detection on Medical Claim Data," in HICSS, 2022.

[2] R. Bauder, R. da Rosa and T. Khoshgoftaar, "Identifying Medicare Provider Fraud with Unsupervised Machine Learning," in IRI, 2018.

[3] WHO, "Global spending on health: a world in transition," World Health Organization, Rep. WHO/HIS/HGF/HFWorkingPaper/19.4, 2019.

[4] Q. Liu and M. Vasarhelyi, "Healthcare fraud detection: A survey and a clustering model incorporating geo-location information," in 29WCARS, Brisbane, Australia, 2013.

[5] I. Kose, M. Gokturk and K. Kilic, "An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance," *Applied Soft Computing*, vol. 36, pp. 283-299, Nov. 2015, doi: 10.1016/j.asoc.2015.07.018.

[6] R. A. Bauder, T. M. Khoshgoftaar, A. Richter and M. Herland., "Predicting medical provider specialties to detect anomalous insurance claims," in ICTAI, 2016.

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

[8] United States Sentencing Commission, "Quick Facts on Health Care Fraud Offenses," United States Sentencing Commission, Rep. FY 2017 through FY 2021 Fraud Team Datafiles, USSCFTFY17-USSCFTFY21.

[9] P. Dora and G. H. Sekharan, "Healthcare insurance fraud detection leveraging big data analytics," *International Journal of Science and Research*, vol. 4, no. 4, pp. 2073-2076, 2015.

[10] D. Thornton, R. M. Mueller, P. Schoutsen and J. V. Hillegersberg, "Predicting healthcare fraud in Medicaid: a multidimensional data model and analysis techniques for fraud detection," *Procedia Technology*, vol. 9, pp. 1252-1264, 2013, doi: 10.1016/j.protcy.2013.12.140.

[11] D. Thornton, M. Brinkhuis, C. Amrit and R. Aly, "Categorizing and describing the types of fraud in healthcare," *Procedia Computer Science*, vol. 64, pp. 713-720, 2015, doi: 10.1016/j.procs.2015.08.594.

[12] S. S. Waghade and A. M. Karandikar, "A comprehensive study of healthcare fraud detection based on machine learning," *International Journal of Applied Engineering Research*, vol. 13, no. 6, pp. 4175-4178, 2018.
[13] R. M. Konijn and W. Kowalczyk, "Finding fraud in health insurance data with two-layer outlier detection approach," in DaWaK, France, 2011.
[14] R. Bauder, T. M. Khoshgoftaar and N. Seliya, "A survey on the state of healthcare upcoding fraud analysis and detection," *Health Services and Outcomes Research Methodology*, vol. 17, pp. 31-55, 2017.
[15] J. Li, K. Y. Huang, J. Jin and J. Shi, "A survey on statistical methods for health care fraud detection," *Health care management science*, vol. 11, pp. 275-287, 2008.
[16] P. Dua and S. Bais, "Supervised learning methods for fraud detection in healthcare insurance," *Machine learning in healthcare informatics*, pp. 261-285, Dec. 2013, doi: 10.1007/978-3-642-40017-9_12.
[17] S. J. Omar, K. Fred and K. K. Swaib, "A state-of-the-art review of machine learning techniques for fraud detection research," in ICSE, Sweden, 2018.
[18] R. Chen, H. Zhang and K. Lin, "A Graph-Based Method for Health Care Joint Fraud Detection," in Proc. ICCPR, 2020.
[19] M. Herland, R. A. Bauder and T. M. Khoshgoftaar, "The effects of class rarity on the evaluation of supervised healthcare fraud detection models," *Journal of Big Data*, vol. 6, no. 1, pp. 1-33, Feb. 2019, doi: 10.1186/s40537-019-0181-8.
[20] R. A. Bauder and T. M. Khoshgoftaar, "The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data," *Health Information Science and Systems*, vol. 6, no. 1, pp. 1-14, Sep. 2018, doi: 10.1007/s13755-018-0051-3.
[21] R. A. Bauder and T. M. Khoshgoftaar, "The detection of medicare fraud using machine learning methods with excluded provider labels," in FLAIRS, 2018.
[22] R. Bauder and T. Khoshgoftaar, "A survey of Medicare data processing and integration for fraud detection," in IRI, 2018.
[23] D. Castro, "Improving health care: why a dose of IT may be just what the doctor ordered," *ITIF Reports*, 2007.
[24] R. Paudel, W. Eberle and D. Talbert. , "Detection of anomalous activity in diabetic patients using graph-based approach," in FLAIRS, 2017.
[25] K. D. Aral, H. A. Güvenir, İ. Sabuncuoğlu and A. R. Akar, "A prescription fraud detection model," *Computer methods and programs in biomedicine*, vol. 106, no. 1, pp. 37-46, Apr. 2012, doi: 10.1016/j.cmpb.2011.09.003.
[26] G. van Capelleveen, M. Poel, R. M. Mueller, D. Thornton and J. van Hillegersberg, "Outlier detection in healthcare fraud: A case study in the Medicaid dental domain," *International journal of accounting information systems*, vol. 21, pp. 18-31, Jun. 2016, doi: 10.1016/j.accinf.2016.04.001.
[27] L. Akoglu, H. Tong and D. Koutra, "Graph based anomaly detection and description: a survey," Data mining and knowledge discovery, vol. 29, pp. 626-688, 2015.
[28] S. Chen and A. Gangopadhyay, "A novel approach to uncover health care frauds through spectral analysis," in ICHI, 2013.
[29] S. L. Wang, H. T. Pai, M. F. Wu, F. Wu and C. L. Li, "The evaluation of trustworthiness to identify health insurance fraud in dentistry,"*Artificial intelligence in medicine*, vol. 75, pp. 40-50, 2017.
[30] D. Huang, D. Mu, L. Yang and X. Cai, "CoDetect: Financial fraud detection with anomaly feature detection," IEEE Access, vol. 6, pp. 19161-19174, 2018
[31] "Healthcare Provider Fraud Detection Analysis," [Online]. Available: https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis.
[32] N. A. Akbar, A. Sunyoto, M. R. Arief and W. Caesarendra, "Improvement of decision tree classifier accuracy for healthcare insurance fraud prediction by using Extreme Gradient Boosting algorithm," in ICIMCIS, 2020.
[33] J. J. Margret and S. Sreenivasan, "Implementation of Data Mining in Medical Fraud Detection," *International Journal of Computer Applications*, vol. 69, no. 5, pp. 1–4, May 2013, doi: 10.5120/11835-7556.
[34] Y. H. Tsai, C.-H. Ko and K. C. Lin, "Using CommonKADS Method to Build Prototype System in Medical Insurance Fraud Detection," *J. Networks*, vol. 9, no. 7, pp. 1798-1802, Jul. 2014, doi: 10.4304/jnw.9.7.1798-1802.
[35] E. A. Duman and Ş. Sağıroğlu., "Heath care fraud detection methods and new approaches," in UBMK, 2017.
[36] H. Cui, Q. Li, H. Li and Z. Yan, "Healthcare fraud detection based on trustworthiness of doctors," in IEEE Trustcom/BigDataSE/ISPA, China, 2016.
[37] T. M. Padmaja, N. Dhulipalla, R. S. Bapi and P. R. Krishna, "Unbalanced data classification using extreme outlier elimination and sampling techniques for fraud detection," in ADCOM, Assam, 2007.
[38] H. Joudaki, A. Rashidian, B. Minaei-Bidgoli, M. Mahmoodi, B. Geraili, M. Nasiri and M. Arab, "Using data mining to detect health care fraud and abuse: a review of literature," *Global journal of health science*, vol. 7, no. 1, Aug. 2014, doi: 10.5539/gjhs.v7n1p194.
[39] L. S. Chen and J. C. Chen, "Using data mining methods to detect medical fraud," in Proc. ICMECG, Jeju Island Republic of Korea, 2020.
[40] K.C. Lin, C.-L. Yeh and S. Y. Huang, "Use of data mining techniques to detect medical fraud in health insurance," *International Journal of Engineering and Technology Innovation*, vol. 2, no. 2, pp. 126-137, 2012.
[41] H. Shin, H. Park, J. Lee and W. C. Jhee, "A scoring model to detect abusive billing patterns in health insurance claims," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7441-7450, Jun. 2012, doi: 10.1016/j.eswa.2012.01.105.
[42] R. Bauder and T. Khoshgoftaar, "Medicare fraud detection using random forest with class imbalanced big data," in IRI, 2018.
[43] J. Hancock and T. M. Khoshgoftaar, "Medicare fraud detection using catboost," in IRI, 2020.
[44] R. A. Bauder, T. M. Khoshgoftaar and A. Napolitano, "Fraud detection with a limited number of known fraudulent medicare providers," in FLAIRS, 2018.
[45] C. Francis, N. Pepper and H. Strong, "Using support vector machines to detect medical fraud and abuse," in EMBC, 2011.
[46] R. A. Bauder and T. M. Khoshgoftaar, "A novel method for fraudulent medicare claims detection from expected payment deviations (application paper)," in IRI, 2016.
[47] S. Lavanya, S. M. Kumar and P. M. Kumar, "Machine Learning Based Approaches for Healthcare Fraud Detection: A Comparative Analysis," *Annals of the Romanian Society for Cell Biology*, pp. 8644-8654, 2021.
[48] H. Farbmacher, L. Löw and M. Spindler, "An explainable attention network for fraud detection in claims management," *Journal of Econometrics*, vol. 228, no. 2, pp. 244-258, Jun. 2022, doi: 10.1016/j.jeconom.2020.05.021
[49] L. Cui, H. Seo, M. Tabar, F. Ma, S. Wang and D. Lee, "Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation," in Proc. ACM SIGKDD, CA, USA, 2020.
[50] M. R. Sumalatha and M. Prabha, "Mediclaim fraud detection and management using predictive analytics," in ICCIKE, 2019.
[51] A. Bayerstadler, L. v. Dijk and F. Winter, "Bayesian multinomial latent variable modeling for fraud and abuse detection in health insurance," *Insurance: Mathematics and Economics*, vol. 71, pp. 244-252, Nov. 2016, doi: 10.1016/j.insmatheco.2016.09.013.
[52] L. K. Branting, F. Reeder, J. Gold and T. Champney, "Graph analytics for healthcare fraud risk estimation," in ASONAM, 2016.
[53] R. A. Bauder and T. M. Khoshgoftaar, "Medicare fraud detection using machine learning methods," in ICMLA, 2017.
[54] A. Bhaskar, S. Pande, R. Malik and A. Khamparia, "An intelligent unsupervised technique for fraud detection in health care systems," *Intelligent Decision Technologies*, vol. 15, no. 1, pp. 127-139, Mar. 2021, doi: 10.3233/idt-200052.
[55] S. Zhou, J. He, H. Yang, D. Chen and R. Zhang, "Big data-driven abnormal behavior detection in healthcare based on association rules," *IEEE Access*, pp. 129002-129011, 2020, doi: 10.1109/access.2020.3009006.
[56] L. F. Carvalho, C. H. Teixeira, W. Meira, M. Ester, O. Carvalho and M. H. Brandao, "Provider-consumer anomaly detection for healthcare systems," in ICHI, 2017.
[57] M. Tang, B. S. U. Mendis, D. W. Murray, Y. Hu and A. Sutinen, "Unsupervised fraud detection in Medicare Australia," in Proc. AusDM, 2011.
[58] M. Kirlidog and C. Asuk, "A fraud detection approach with data mining in health insurance," *Procedia-Social and Behavioral Sciences*, vol. 62, pp. 989-994, Oct. 2012, doi: 10.1016/j.sbspro.2012.09.168.
[59] D. Lasaga and P. Santhana, "Deep learning to detect medical treatment fraud," in KDD 2017 Workshop on Anomaly Detection in Finance, PMLR, 2018.
[60] R. M. Musal, "Two models to investigate Medicare fraud within unsupervised databases," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8628-8633, Dec. 2010, doi: 10.1016/j.eswa.2010.06.095.
[61] R. A. Bauder and T. M. Khoshgoftaar, "A probabilistic programming approach for outlier detection in healthcare claims," in ICMLA, 2016.
[62] A. Gangopadhyay and S. Chen, "Health care fraud detection with community detection algorithms," in SMARTCOMP, 2016.
[63] G. C. Capelleveen, "Outlier based predictors for health insurance fraud detection within US Medicaid," M.S. thesis, SMG, UT, 2013.
[64] D. Thornton, G. v. Capelleveen, M. Poel, J. v. Hillegersberg and R. M. Mueller, "Outlier-based Health Insurance Fraud Detection for US Medicaid Data," in ICEIS, 2014.

[65] H. Peng and M. You, "The health care fraud detection using the pharmacopoeia spectrum tree and neural network analytic contribution hierarchy process," in IEEE Trustcom/BigDataSE/ISPA, 2016.

[66] T. Ekina, F. Leva, F. Ruggeri and R. Soyer, "Application of Bayesian methods in detection of healthcare fraud," *Chemical Engineering Transaction*, vol. 33, 2013.

[67] S. Zhu, Y. Wang and Y. Wu, "Health care fraud detection using nonnegative matrix factorization," in ICCSE, 2011.

[68] C. Sun, Z. Yan, Q. Li, Y. Zheng, X. Lu and L. Cui, "Abnormal group-based joint medical fraud detection," *IEEE Access*, vol. 7, pp. 13589-13596, 2019, doi: 10.1109/access.2018.2887119.

[69] J. Seo and O. Mendelevitch, "Identifying frauds and anomalies in medicare-B dataset," in EMBC, 2017.

[70] J. Liu, E. Bier, A. Wilson, J. A. Guerra-Gomez, T. Honda, K. Sricharan, L. Gilpin and D. Davies, "Graph analysis for detecting fraud, waste, and abuse in healthcare data," *AI Magazine*, vol. 37, no. 2, pp. 33-46, Jun. 2016, doi: 10.1609/aimag.v37i2.2630.

[71] L. Copeland, D. Edberg, A. K. Panorska and J. Wendel, "Applying business intelligence concepts to Medicaid claim fraud detection," *Journal of Information Systems Applied Research*, vol. 5, no. 1, p. 51, 2012.

[72] V. Rawte and G. Anuradha, "Fraud detection in health insurance using data mining techniques," in ICCICT, 2015.

[73] C. Ngufor and J. Wojtusiak, "Unsupervised labeling of data for supervised learning and its application to medical claims prediction," *Computer Science*, vol. 14, no. 2, p. 191, 2013, doi: 10.7494/csci.2012.14.2.191.

[74] C. Sun, Q. Li, H. Li, Y. Shi, S. Zhang and W. Guo, "Patient cluster divergence based healthcare insurance fraudster detection," *IEEE Access*, vol. 7, pp. 14162-14170, 2019, doi: 10.1109/access.2018.2886680.

[75] M. Irum, S. A. Khan and H. U. Rahman, "Sequence mining and prediction-based healthcare fraud detection methodology," *IEEE Access*, vol. 8, pp. 143256-143273, 2020, doi: 10.1109/access.2020.3013962.

[76] C. P. Killen and C. Kjaer, "Understanding project interdependencies: The role of visual representation, culture and process," *International Journal of Project Management*, vol. 30, no. 5, pp. 554-566, Jul. 2012, doi: 10.1016/j.ijproman.2012.01.018.

[77] A. Abdallah, M. A. Maarof and A. Zainal, "Fraud detection system: A survey," *Journal of Network and Computer Applications*, vol. 68, pp. 90-113, Jun. 2016, doi: 10.1016/j.jnca.2016.04.007.

[78] A. Popov, "Feature engineering methods," in *Advanced Methods in Biomedical Signal Processing and Analysis,* Academic Press, 2023, pp. 1-29.

[79] J. M. Johnson and T. M. Khoshgoftaar, "Medicare fraud detection using neural networks," *Journal of Big Data*, vol. 6, no. 1, pp. 1-35, Jul. 2019, doi: 10.1186/s40537-019-0225-0.

[80] M. Herland, T. M. Khoshgoftaar and R. A. Bauder, "Big data fraud detection using multiple medicare data sources," *Journal of Big Data*, vol. 5, no. 1, pp. 1-21, Sep. 2018, doi: 10.1186/s40537-018-0138-3.

[81] J. M. Johnson and T. M. Khoshgoftaar, "Medical provider embeddings for healthcare fraud detection," *SN Computer Science*, vol. 2, no. 4, May 2021, doi: 10.1007/s42979-021-00656-y.

[82] M. Köppen, "The curse of dimensionality," in WSC5, 2000.

[83] S. Arora, Y. Liang and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," in ICLR, 2017.

[84] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski and S. Ananiadou, "Distributional semantics resources for biomedical text processing," in LBM, 2013.

[85] R. A. Sowah, M. Kuuboore, A. Ofoli, S. Kwofie, L. Asiedu, K. M. Koumadi and K. O. Apeadu, "Decision support system (DSS) for fraud detection in health insurance claims using genetic support vector machines (GSVMs)," *Journal of Engineering*, vol. 2019, pp. 1–19, Sep. 2019, doi: 10.1155/2019/1432597.

[86] R. A. Bauder, T. M. Khoshgoftaar and T. Hasanin, "Data sampling approaches with severely imbalanced big data for medicare fraud detection," in ICTAI, 2018.

[87] M. Xu, "Understanding graph embedding methods and their applications," *SIAM Review*, vol. 63, no. 4, pp. 825-853, 2021.

[88] W. Fan, "Big graphs: challenges and opportunities," *Proceedings of the VLDB Endowment*, vol. 15, no. 12, pp. 3782-3797, Aug. 2022, doi: 10.14778/3554821.3554899.

[89] P. Cui, X. Wang, J. Pei and W. Zhu, "A survey on network embedding," *IEEE transactions on knowledge and data engineering*, vol. 31, no. 5, pp. 833-852, 2018.

[90] H. Cai, V. W. Zheng and K. C.-C. Chang, "A comprehensive survey of graph embedding: Problems, techniques, and applications," *IEEE transactions on knowledge and data engineering*, vol. 30, no. 9, pp. 1616-1637, 2018.

[91] Y. Bengio, A. Courville and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798-1828, 2013.

[92] A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski and A. J. Smola, "Distributed large-scale natural graph factorization," in Proc. WWW '13, pp. 37-48, 2013.

[93] M. Ou, P. Cui, J. Pei, Z. Zhang and W. Zhu, "Asymmetric transitivity preserving graph embedding," in Proceedings of the 22th ACM SIGKDD international conference on Knowledge discovery and data mining, 2016.

[94] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Advances in neural information processing systems*, vol. 14, 2001.

[95] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[96] B. Perozzi, R. Al-Rfou and S. Skiena, "Deepwalk: Online learning of social representations," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014.

[97] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016.

[98] D. Wang, P. Cui and W. Zhu, "Structural deep network embedding," in Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016.

[99] S. Cao, W. Lu and Q. Xu, "Deep neural networks for learning graph representations," in Proceedings of the AAAI conference on artificial intelligence, 2016.

[100] P. Goyal and E. Ferrara, "Graph embedding techniques, applications, and performance: A survey," *Knowledge-Based Systems*, vol. 151, pp. 78-94, 2018.

[101] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.

[102] W. Hamilton, Z. Ying and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[103] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio and Y. Bengio, "Graph attention networks," *stat*, vol. 1050, no. 20, pp. 10-48550, 2017.

[104] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[105] Y. Ren and J. Zhang, "HGAT: hierarchical graph attention network for fake news detection," *arXiv preprint arXiv:2002.04397*, 2020.

[106] J. Choi, J. Park, W. Kim, J.-H. Park, Y. Suh and M. Sung, "PU GNN: Chargeback Fraud Detection in P2E MMORPGs via Graph Attention Networks with Imbalanced PU Labels," *arXiv preprint arXiv:2211.08604*, 2022.

[107] J. Tao, H. Wang and T. Xiong, "Selective graph attention networks for account takeover detection," in ICDMW, 2018.

[108] C. Liu, L. Sun, X. Ao, J. Feng, Q. He and H. Yang, "Intention-aware heterogeneous graph attention networks for fraud transactions detection," in the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021.

[109] [Online]. Available: https://www.cdc.gov/nchs/icd/icd9cm.htm.

[110] [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

[111] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

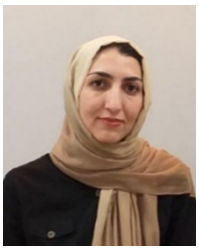
[112] S. Yun, M. Jeong, R. Kim, J. Kang and H. J. Kim, "Graph transformer networks," *Advances in neural information processing systems*, vol. 32, 2019.

[113] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, pp. 197-227, 2016.

[114] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[115] D. W. Hosmer Jr, S. Lemeshow and R. X. Sturdivant, Applied logistic regression, vol. 398, John Wiley & Sons, 2013.

[116] "Medicare Physician & Other Practitioners," Centers for Medicare & Medicaid Services, [Online]. Available: https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners.



SHAHLA MARDANI received the M.S degree in information technology engineering from the University of Amirkabir, Tehran, Iran, in 2013. She is a PhD candidate in the field of information technology in the School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran. She researches collusive fraud detection in the healthcare domain in her thesis and her current research interests include fraud detection, graph analytics, and machine learning.

HADI MORADI (SM'88) received the B.Sc. degree in electrical engineering from the School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran, in 1988, and the Ph.D. degree in computer engineering from the University of Southern California, Los Angeles, CA, USA, in 2012. He is currently a Professor with the School of Electrical and Computer Engineering, University of Tehran. His current research interests include data analysis and pattern recognition in cognitive games, intelligent cognitive screening and rehabilitation, and robotics.