Expert Systems With Applications 215 (2023) 119259

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

WMTDBC: An unsupervised multivariate analysis model for fraud detection in health insurance claims

Lavanya Settipalli, G.R. Gangadharan *

Department of Computer Applications, National Institute of Technology, Tiruchirappalli, India

ARTICLE INFO

Keywords:
Healthcare fraud detection
Multivariate analysis
Weighted MultiTree
Density based clustering

ABSTRACT

Fraud is an aggravating problem in the health insurance system, causing a substantial increase in the cost of medical services. Many models have been developed using data mining or machine learning techniques to lessen the impact of fraud on healthcare system. However, achieving good accuracy is still challenging as the claims data is multivariate with multiple class overlappings. In this paper, we propose a novel approach of unsupervised multivariate analysis for healthcare claims submitted by the providers. Our proposed model analyzes multivariate categorical data and continuous data in two stages to observe providers' behaviors. The first stage constructs Weighted MultiTree (WMT) for categorical data to analyze similarity among provider profiles, the relation among profiles, and rendered services to identify false services. The second stage detects false claims by developing a univariate fraud detection model using different Density Based Clustering (DBC) techniques on continuous data of claims such as service counts and service charges. The performance of our proposed WMTDBC is measured by conducting experiments on the claims within various medical specialties or provider types of CMS part B program. Our empirical results evidence that the detection performance is enhanced with our WMTDBC approach when compared with the state-of-the-art models.

1. Introduction

A health insurance policy is an agreement between a health insurance company and a beneficiary to provide financial coverage for medical expenses. These policies provide either complete reimbursement of medical expenses or free-of-cost treatments to the beneficiary. The ever increasing costs of healthcare tend the patients, service providers such as doctors or healthcare organizations, and health insurance providers to commit fraud. Many sources reveal the facts about fraud amount estimations and their impact on the rise of healthcare costs. As per the National Health Care Anti-Fraud Association (NHCAA), the financial loss due to healthcare fraud is about USD 68 billion each year (King, 2014; Simborg, 2008). A conservative estimate is 3% of total healthcare expenditures, while some government and law enforcement agencies estimating the loss as high as 10% of the annual health outlay, which could mean approximately USD 300 billion.[1] According to the statistics of the Department of Health and Human Services (HHS), the total US medicare spending raised from USD 471 billion to USD 798 billion in the span of a decade from 2009 to 2019.[2] Moreover, the taxation of the Affordable Care Act (ACA) mentioned the growth rate of healthcare fraud as 6.2% per annum between 2015 and 2021.[3] Although the extent of healthcare fraud against the Australian private health insurance funds is unknown, it would be in the range of half a percent to one and half a percent or it would be around seven percent or higher (Flynn, 2015). In India, it is estimated that the number of fraudulent claims in the healthcare industry is approximately 15% of total claims.[4] These facts indicating that the healthcare industries are under pressure to control the increase in healthcare spending growth rate by detecting and reducing wasteful spendings. Hence this study focuses on detecting healthcare fraud especially committed by the providers since they cause a great loss to the healthcare compared to the other possible frauds such as subscriber fraud or carrier fraud (Pflaum & Rivers, 1991).

Provider fraud is an intentional act by healthcare providers to get unlawful benefits from the insurance companies causes an unnecessary increase in the costs of services and also lessens the quality of health

* Corresponding author.
E-mail addresses: lavanya.sp86@gmail.com (L. Settipalli), geeyaar@gmail.com (G.R. Gangadharan).

[1] https://www.bcbsm.com/health-care-fraud/fraud-statistics.html.
[2] https://www.hhs.gov/sites/default/files/fy-2019-budget-in-brief.pdf.
[3] https://weaver.com/blog/affordable-care-act-and-health-care-fraud.
[4] https://www.https://www.insuranceinstituteofindia.com/c/document_library/get_file?uuid=e4632c21-da80-494c-9264-395283e3e4c0&groupId=16940.

https://doi.org/10.1016/j.eswa.2022.119259
Received 30 August 2021; Received in revised form 29 May 2022; Accepted 9 November 2022
Available online 19 November 2022

services. There are different ways that a service provider can commit fraud such as billing for the services or goods that are not rendered, performing medically unnecessary services, prescribing unnecessary medicines or services, submitting excessive charges for rendered services, and malpractices, etc. Hence, provider fraud has become difficult to assess, and most healthcare systems have no consistent methods to measure them. Earlier, the traditional fraud detection system involved a medical expert group on behalf of insurance carriers to assess the frauds in the submitted claims of the providers. However, the traditional approach of manually auditing huge amounts of claims delays the process of fraud detection (Ashtiani & Raahemi, 2021). Hence, to mitigate the problem of manual auditing systems, an era of using machine learning techniques and deep learning techniques (Ashtiani & Raahemi, 2021; Ozbayoglu, Gudelek, & Sezer, 2020) has been started for automating the auditing process. Several potential researches have come towards automating the fraud detection process using machine learning techniques not only in healthcare but also in automobile insurance (Dhieb, Ghazzai, Besbes, & Massoud, 2020), in banking and credit card systems (Boutaher, Elomri, Abghour, Moussaid, & Rida, 2020; Carrasco & Sicilia-Urbán, 2020; Lucas et al., 2020). In healthcare, the major challenge is the availability of claims data to develop models as the organizations avoid storing healthcare records publicly due to the risk of vulnerability. Although there exist advanced techniques for secure storage of these digitized records or Electronic Health Records (EHRs) (Chelladurai & Pandian, 2021; Mubarakali, Bose, Srinivasan, Elsir, & Elsier, 2019), the fraudsters follow tricky ways to commit fraud. Most of the fraud detection systems developed supervised learning approaches based on audited records with the involvement of healthcare experts to decide the set of core features and train the fraud detection system. However, the availability of audited records in healthcare is very limited and the involvement of medical experts delays and endures the development of a fraud detection system. Hence, developing a fraud detection model using unsupervised learning techniques that do not need any prior knowledge about anomalies or fraudulent events would be a better way to combat the issue of the availability of labeled data.

The main aim of the proposed model is to identify the fraudulent activities of the providers' such as rendering services that are irrelevant to the profile, furnishing unnecessary services, and the excessive charges claimed by the providers than the actual cost of the rendered service. Fraudulent activities of a provider can be analyzed when comparing them among the claims of providers belong to the same or similar community. A similar community of claims can be formed based on the similarity in their profiles and their behaviors (Jiang et al., 2019). The provider behavior includes the rendered services, the number of services rendered per day, and the amounts claiming for rendered services. However, analyzing and grouping a huge number of claims with a wide variety of provider profiles, identifying the relationship between provider profiles and the rendered services, analyzing the service counts and claimed amounts of similar claims, altogether is a complex task. Hence we divide the process of identifying the fraudulent activities into two stages to reduce the complexity in analyzing the claims. The two stages trail a hierarchical order in which the first stage analyzes categorical information to group the claims with similar profiles, and analyzes the relationship between the provider profiles and services by constructing Weighted MultiTree (WMT). The WMT is an unsupervised graph based approach that can be constructed for the providers' claims by considering profiles and service details as nodes and connect them based on the relationship as defined in the claims data. The edges connecting the profiles and the service details are assigned with weights based on the frequency of their occurrence. Once the WMT is constructed, the services that are connected to profiles with lesser weights than the desired threshold will be considered as the services irrelevant to the profiles. Then the second stage of the model analyzes the number of services furnished by a provider and the claimed amounts among a similar group of claims to identify unnecessary services and excessive charges for the associated rendered services. The second stage applies different density based clustering techniques on each group of similar claims to detect the deviations in the service counts and claimed amounts. Density based clustering techniques such as DBSCAN (Sander, Ester, Kriegel, & Xu, 1998) and its other three variants OPTICS (Ankerst, Breunig, Kriegel, & Sander, 1999), DENCLUE (Hinneburg, Keim, et al., 1998), and GDBSCAN (Ester, Kriegel, Sander, Xu, et al., 1996) are applied to each similar group of claims to compare the service counts and submitted amounts with the competent providers. We evaluate the performance of our approach by analyzing the fraud cases detected from the claims submitted by the providers in the financial year 2018 for rendering services under the US medicare CMS part B[5] healthcare program.

The remainder of this paper is organized as follows. Section 2 discusses the research works related to healthcare fraud detection. Section 3 describes the construction of Weighted MultiTree based on the provider profiles and services rendered and density based clustering on the groups of similar claims. Section 4 presents the results and discussion about the performance of our model. Section 5 contains the concluding remarks of our work.

2. Related work

Several models have been developed by researchers to make the healthcare fraud detection system automated using supervised, semi-supervised and unsupervised machine learning techniques. Chandola, Sukumar, and Schryver (2013) analyzed the key challenges in detecting fraudulent healthcare claims and analyzed the advancement in knowledge discovery approaches for fraud detection. Joudaki et al. (2015) presented supervised and unsupervised data mining techniques for detecting healthcare fraud and abuse. The review has mostly focused on the studies that developed based on algorithmic data mining. Ahmad, Qamar, and Rizvi (2015) reviewed the state-of-the-art data mining techniques such as classification, clustering, association, and regression which are used to develop the healthcare fraud detection system. It also highlighted the applications, challenges, and advancements of data mining approaches applicable to the healthcare domain. Bauder, Khoshgoftaar, and Seliya (2017) reviewed the state-of-the-art models of upcoding analysis where upcoding is defined as the exaggeration of the rendered service details by providers to a more expensive procedure code, in order to get additional reimbursement. This review discussed the supervised, unsupervised, and hybrid models for the upcoding analysis of healthcare claims that are classified under several case-mix systems or coding techniques such as DRG (Diagnosis Related Group), ICD (International Classification of Diseases), CPT (Current Procedural Terminology), HCPCS (Healthcare Common Procedure Coding System), RUG (Resource Utilization Groups), and MDC (Major Diagnostic Category). The aforementioned papers reviewed the models to detect frauds by physicians submitting excessive charges for the rendered services by comparing the costs of services among similar communities. Whereas, in this section, we presented a review of some models in two different aspects such as models developed based on supervised learning approaches and models developed based on unsupervised learning approaches.

2.1. Supervised learning approaches

There are many discretionary approaches to healthcare frauds detection using supervised learning techniques. He, Graco, and Yao (1998), He, Wang, Graco, and Hawkins (1997) and He, Hawkins, Graco, and Yao (2000) developed supervised learning based models for health insurance fraud detection. The method developed by He et al. (1997)

[5] https://www.cms.gov/research-statistics-data-systems/medicare-provider-utilization-and-payment-data/medicare-provider-utilization-and-payment-data-physician-and-other-supplier/physician-and-other-supplier-data-cy-2018.

used Self Organizing Maps (SOM) and neural networks with an error backpropagation algorithm to group similar providers and identify the abnormal profiles of the providers. The methods developed in He et al. (1998, 2000) calculated the optimized weights of different practices of providers using the K-nearest neighbor (KNN) algorithm and identified incorrect practices by comparing them with the normal practice profiles. Ormerod, Morley, Ball, Langley, and Spenser (2003) proposed a supervised Bayesian network for detecting frauds in health insurance claims. The weights are calculated and assigned as an automatic indicator for each fraud type by observing the patterns of repeated frauds. Ortega, Figueroa, and Ruz (2006) proposed a supervised multilayer perceptron neural network (MLP) for the detection of frauds. They have used multilayer perception neural networks to analyze the hidden patterns of the nonlinear dependencies in the provider profiles. The usage of the fraud indicators defined by this approach is limited to the claims provided by a private health insurance company from Chile. Yang and Hwang (2006) proposed a data mining based framework for adaptable and extensible detection of healthcare fraud. The model analyzed the structure patterns of clinical instances compromising a set of activities performed by the medical staff. The frequently structured patterns are considered as the features and a subset of feature space which shows the high impact in reducing running time and increasing accuracy is taken through the Markov blanket filter approach.

Kumar, Ghani, and Mei (2010) proposed a fraud detection method for health insurance claims using supervised support vector machines (SVM). The abnormalities in insurance claims by the providers are predicted by converting the categorical features to binary features. This conversion extended the ability of SVM to handle categorical data of healthcare claims with a large number of features. Herland, Khoshgoftaar, and Bauder (2018) applied Logistic Regression (LR), Gradient Boosted Trees (GBT), and Random Forests (RF) to analyze and compare the performance of those approaches in detecting the fraudulent behaviors of providers using the combined claims data of CMS part B, List of Excluded Individuals/Entities (LEIE),[6] and Durable Medical Equipment, Prosthetics/Orthotics & Supplies (DMEPOS).[7] However, they have not analyzed the homogeneity among the claims before applying the selected learning techniques. In their other work (Herland, Bauder, & Khoshgoftaar, 2020), they detected the fraudulent behaviors of physicians by predicting the specialty based on the provider type and the number of services rendered. They have performed specialty grouping, class removal, and class isolation for analyzing homogeneous claims and devised two methods Combining Office and Facility (COF) and Separating Office and Facility (SOF) while class grouping. Matloob, Khan, and Rahman (2020) provides a better insight into how to improve patient management and treatment procedures by generating and analyzing the sequence of services availed by the patients for a specialty. The proposed model used the Bayes rule for populating frequent and rare sequences and developed two modules: The sequence rule engine analyzes the frequent sequence patterns using the Prefixspan pattern sequence mining algorithm and the Prediction based engine analyzes the rare pattern sequences using the Compact Prediction Tree to detect fraud cases

Zhou, He, Yang, Chen, and Zhang (2020) proposed a big data-driven healthcare fraud detection framework in order to analyze the huge number of claims with less time complexity. The Association rule mining is applied on claims based on the frequent pattern mining on healthcare claims using distributed computing. Johnson and Khoshgoftaar (2021) converted the provider types and their descriptions such as drug prescriptions and provided treatments into vector space using embedding techniques such as GloVe, Med-W2V, HcpcsVec, and RxVec. The embedded spaces of similar provider types are then compared to analyze for the deviations. They have applied different classifiers including LR, RF, Gradient Boosted Tree (GBT), and MultiLayer Perceptron (MLP) on embedded space in order to evaluate the performance of different embedding techniques. Haque and Tozal (2021) proposed a novel healthcare fraud detection model by determining the claims as a Mixture of latent Clinical Concepts (MCC) in a latent space using probabilistic topic modeling. The hierarchical relationship structure of diagnosis and procedure codes are drawn based on the example claims. Then MCC which is a representation learning process extracts features where the clinical concepts have a high frequency. The features extracted from the MCC are fed to LSTM (Long–Short Term Memory) for representing the claim as a sequence of dependent concepts to outlier irrelevant concepts. This model also represented another variant using Robust Principal Component Analysis (RPCA) to decompose low-rank and sparse vector representation of claims. Hancock and Khoshgoftaar (2021) compared the performance of various boosting approaches including CatBoost, XGBoost, Gradient Boost Decision Trees (GBDT) in detecting fraudulent claims by applying them on the Medicare dataset. This model considered the categorical information of providers including HCPCS code and the provider state as input features to classify among the fraudulent and non-fraudulent claims.

2.2. Unsupervised learning approaches

Due to the diversities in healthcare data and difficulty in obtaining labeled data from medical field relative domains (Štefan & Bajec, 2008), applying supervised learning in detecting frauds is limited, and hence researchers are focusing on unsupervised learning techniques to develop fraud detection systems. Yamanishi, Takeuchi, Williams, and Milne (2004) proposed a novel SmartSifter algorithm, unsupervised learning to detect outliers in Australian medical insurance pathology transaction data that can handle both continuous and categorical variables. SmartSifter algorithm calculates the scores to the claims according to the similarities among the variables and identifies claims with high scores as outliers. Luo and Gallagher (2010) proposed an algorithm to identify the DRG upcoding by partitioning the claims of the healthcare system that follows DRG coding for processing procedures. The algorithm partitioned the claims into homogeneous subgroups using decision tree learning in the absence of historical audit data. Johnson and Nagarur (2016) developed a fraud risk detection model at six different stages by observing the abnormalities in the provider claims at different stages. In the initial three stages, they performed provider profiling by observing the similarity among the provider profiles using Mahalanobis distance and they used density based calculation for provider profiling. In the fourth stage, they quantified the fraud risk for the claims by integrating the information of abnormalities from the first three stages. Then they determined the risk threshold value using decision trees in the fifth stage and compared the risk values and risk threshold to assess the risk of fraud in the claims. However, calculating the similarities using distance based approaches leaves the intermediate results in finding variations among provider profiles.

Bayerstadler, van Dijk, and Winter (2016) developed a framework called the multinomial Bayesian latent variable model in order to observe the fraud behaviors of medical providers and insured members. For this purpose, the habitual patterns of providers have been observed to identify the systematic deviations in the claims of providers. The deviations are measured based on the estimations using Markov Chain Monte Carlo (MCMC) algorithm. van Capelleveen, Poel, Mueller, Thornton, and van Hillegersberg (2016) presented unsupervised outlier techniques to analyze the healthcare claims after payments and designed decision support tools that enable the domain experts to take decisions on fraudulent behaviors rapidly. Bauder and Khoshgoftaar (2016) observed the probability distribution of healthcare claims submitted by the providers to discern the deviated behaviors of the

[6] https://oig.hhs.gov/exclusions/exclusions_list.asp.

[7] https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/DME2018.

providers. This method incorporated a Bayesian inference model with credible intervals to enhance the confidence levels in detecting fraudulent claims. In the later year, Bauder and Khoshgoftaar (2017) developed a two-folded multivariate model to analyze multivariate claims data. Multivariate Adaptive Regression Splines (MARS) model is used in the first step for producing studentized residuals and the outcome residuals are used as input to develop a univariate outlier fraud detection model in the second step. The outlier claims in the second step are analyzed based on full Bayesian Inference (BI), using probabilistic programming. Sowah et al. (2019) proposed Genetic support vector machines (GSVMs), a hybrid learning method that is incorporated with both supervised and unsupervised learning. Another unsupervised approach developed by Naidoo and Marivate (2020) analyzed anomalies based on the labels generated by Generative Adversarial Networks (GAN), unlike the other outlier detection approaches that generate normal profiles to flag values that significantly deviate from the acceptable patterns. The labeling of claims to normal or anomalous was done using the loss function of GAN which can also be called an anomaly score.

Most of the models developed for detecting frauds in health insurance claims used supervised learning approaches such as decision trees and neural networks. The models developed using decision trees need to impose many rules from the huge number of claims which is highly complex. Healthcare claims vary from provider to provider based on the patterns in their practices and profiles. The claims of a provider need to be compared with another provider who has similar profiles and patterns of services. Most of the unsupervised models analyzed healthcare claims without subgrouping the homogeneous claims causing false alarms in detecting fraudulent health insurance claims. Some existing models subgrouped similar claims using clustering techniques and neural networks. However, they tend to generate intermediate results in achieving a clear separation among different groups of claims. The intermediate results occur since different groups of claims contain most of the profile information and service patterns in common. Our proposed Weighted MultiTree (WMT) model is an unsupervised approach that does not require any labeled data or the involvement of medical experts. Our model is inspired by the concept of MultiTree (Moshagen, 2010) which is a searchable database of hypotheses on the relationships among different data variables. In combinatorics and order-theoretic mathematics, a MultiTree (MT) is a Directed Acyclic Graph (DAG) in which the set of vertices is reachable from any vertex. The WMT approach introduced in our model analyzes the frequent patterns of services by the providers with specific profile information in order to observe the relationship between the profile of providers and the rendered services. The WMT also allows different providers to share common information among them. As the proposed WMT is a graph based approach having unambiguous nature, the clear deviation between different groups of claims can be attained without any intermediate results. The unsupervised multivariate analysis model developed in this study has the following advantages over the existing models of detecting frauds in health insurance claims:

1. The model developed in our study has used unsupervised learning techniques in each stage which does not require any audited or labeled data.
2. The proposed model is able to perform multivariate analysis on both categorical and continuous types of data that generally exist in health insurance claims.
3. The graph based approach proposed in this model using MT for grouping similar claims will not be affected by any intermediate results as MTs are unambiguous in defining the relationship between the nodes.
4. The profiles of different providers have most of the data in common. As our proposed model is developed based on the property of MT, which defines that a set of nodes be reachable from any node, the WMT constructed for providers details will be with less number of nodes since it shares the common nodes among different providers. Hence it is possible to analyze the relation among the providers and their details with reduced construction cost by avoiding the creation of redundant nodes.
5. The density based clustering techniques used in our approach to identifying unnecessary services and excessive charges can cluster similar claims even the services count and claimed amounts distribute in any shape.

3. An unsupervised multivariate analysis for health insurance fraud detection

The unsupervised model developed in this study has two stages: The first stage constructs Weighted MultiTree (WMT) for grouping the claims of the providers with similar provider information and service details. The WMT also analyzes the relationship among the profiles and the services to identify the services rendered by the providers that are irrelevant to their profile. The irrelevant services identified by the WMT will be considered as fraudulent services of the providers. The second stage analyzes the services count and claimed amounts submitted by the providers for the rendered services by comparing them with that of the competent providers having similar profiles and service details. The density based clustering techniques are used for this purpose because of their property of effectively defining the correlation among the data points. The density based clustering techniques can also cluster the services count and claimed amounts irrespective of their distribution, unlike other clustering techniques which cluster the data points in circular shape based on the centroid. The framework of our model is as shown in Fig. 1.

3.1. Weighted MultiTree construction

However, some common provider profile details and service details among the providers with different provider types exist. The WMT allows cross connections among different details of the provider types. The frequency of occurrence of an edge E between the root node and the detail or between two details is assigned as a frequency F for that edge. That means, if a new detail has appeared for a provider type, then the weight of the edge that connects the detail to the MultiTree is 1. If the same detail is appearing repeatedly, then F is the frequency of occurrence of the same detail under that provider type. For example, we consider the claims dataset of five providers with demographics and service details as $\{P_1, a_1, b_1, c_1\}$, $\{P_1, a_2, b_2, c_3\}$, $\{P_2, a_1, b_1, c_1\}$, $\{P_1, a_1, b_1, c_1\}$, and $\{P_2, a_2, b_1, c_2\}$ where P_1 and P_2 are provider types and c_1, c_2, and c_3 are the services rendered (See Fig. 2). In the claims of the providers, three providers belong to the provider type P_1 and two providers belong to the provider type P_2.

A MultiTree (MT) is a Directed Acyclic Graph (DAG) in which a set of nodes can be reachable from any node. As MT is unambiguous, MT is used in our model to extract the connections among the profile of a provider and the rendered service details. We have used the concept of assigning weights to the MT as a deciding factor to know the services that are irrelevant to the provider profiles. As we aim to analyze the claims within each medical specialty or provider type, the WMT is constructed by considering each unique provider type as a root node and connected the remaining profile details to the root node as appear in the claims. The services rendered by the providers with corresponding profiles are appended as the leaf nodes of the MultiTree. Let us assume that $G = (V; E; F)$ is a MultiTree, where V is the detail, E is the edge, and F is the frequency value of the edge E. Initially, each distinct provider type is taken as a root node, and the remaining profile details of providers as intermediate nodes, are connected through the edges to the corresponding root node.

The WMT for the said claims of the providers can be constructed as shown in Fig. 3. We consider two provider types P_1 and P_2 as root nodes in the graph. The other profile details are taken as intermediate

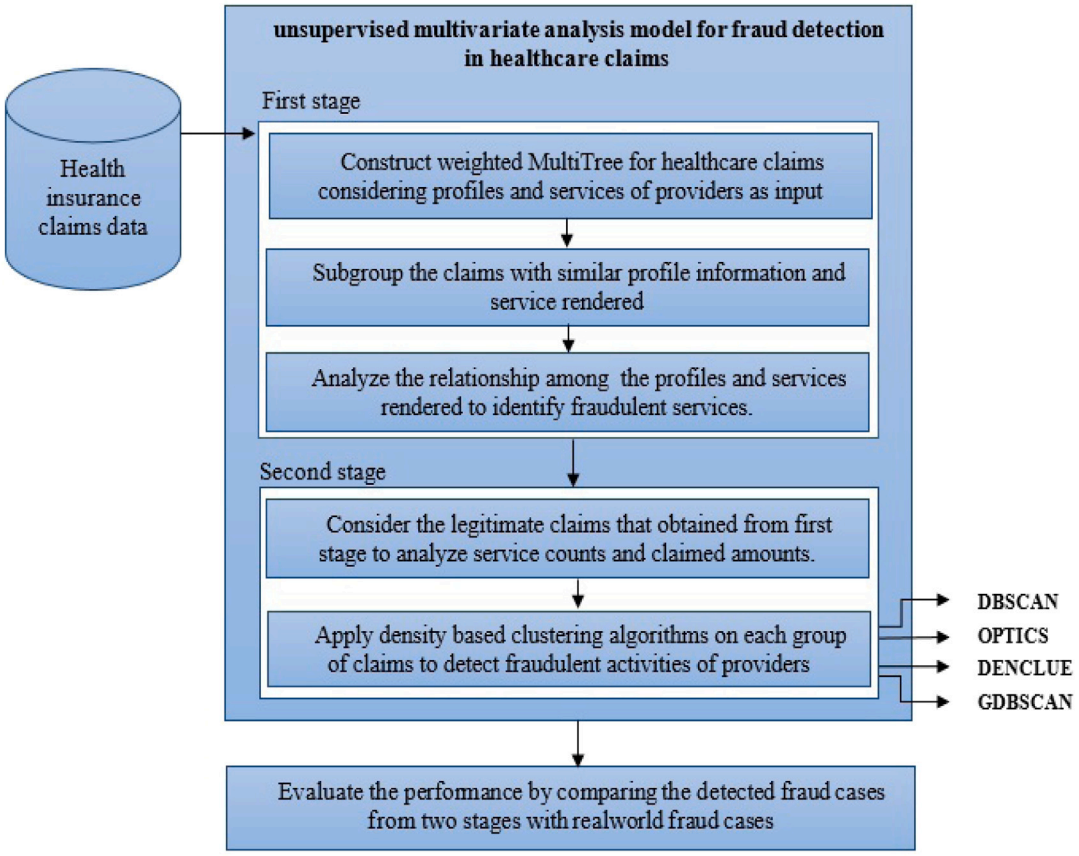


Fig. 1. A framework of a unsupervised multivariate analysis model for detecting fraud in health insurance claims.

Service details

	Provider Type	Dem_info_2	Serv_detail_1	Service furnished
Provider_1	P_1	a_1	b_1	c_1
Provider_2	P_1	a_2	b_2	c_3
Provider_3	P_2	a_1	b_1	c_1
Provider_4	P_1	a_1	b_1	c_1
Provider_5	P_2	a_2	b_1	c_2

Demographic information

Fig. 2. Sample claims dataset with assumed details.

nodes that are connected to the root nodes as per the relation among them mentioned in the dataset. The services rendered are taken as the leaf node and connected to the associated paths. However, there is no particular hierarchy to be followed when constructing the WMT. All the claim details are to be taken in the same order as the first claim details were taken. Since all the details are categorical information, the order of the remaining details will not affect subgrouping the similar claims. From Fig. 3, it can be observed that the provider types of two different providers sharing the nodes of common provider details through the cross connections in WMT without a new node for each provider detail. This feasibility of WMT drastically reduces the cost of constructing a network among providers and their details and provides high reliability in analyzing the relationship among them. It can also be observed that the edge connections are assigned with the weights based on the frequency of co-occurrence of the provider details (which are connecting by the edge) in the dataset. The weight of an edge

between two details P_1 and a_1 is 2 and between the details a_1 and b_1 is 3 i.e. (2 + 1). Algorithm 1 describes the detailed steps for constructing a WMT for providers from health insurance claims data.

Algorithm 1. Constructing Weighted MultiTree for provider details

Input: Health insurance claim dataset D with provider demographics and services details.
Output: MultiTree for provider details

```
define columns from dataset D for provider details
for each record R in D do
  if (new provider type)
    define visited as empty list and create provider type as root node
    current = first column detail element
    connect an edge between root node to current
    initialize edge count to 1
    while (current ≠ 0) do
      append current to visited list
      next = current. next detail element
      if (next ≠ 0)
        connect an edge between current to next
        initialize edge count to 1
      end
      current = next
    end
  end
  else if (existing provider type)
    current = first column detail element
    if (current is in visited)
      increment edge count between root node to current
    else
      connect an edge between root node to current
      initialize edge count to 1
      append current to visited list
    end
    while (current ≠ 0) do
      next = current . next detail element
      if (next ≠ 0 && not in visited)
        connect an edge between current to next
        initialize edge count to 1
        append current to visited list
      end
      else if (next ≠ 0 && is in visited)
        increment count of edge between current to next node
        current = next
      end
    end
  end
end
return MultiTree
```

The complexity of Algorithm 1 can be analyzed as the number of details visited to construct the WMT. Let us assume that the number of columns considered for profile details as C and the total number of unique claim records as R. Then the number of details to be visited are CR. Note that even the repeated details are formed as a single node, each and every detail should be visited for constructing MultiTree. And hence, the complexity is derived based on the total number of details. The complexity for assigning the frequency to each path also consumes the same as constructing the constructing MultiTree. Therefore, the complexity of constructing a MultiTree and assigning weights to each path will be T(CR + CR) i.e. T(CR).

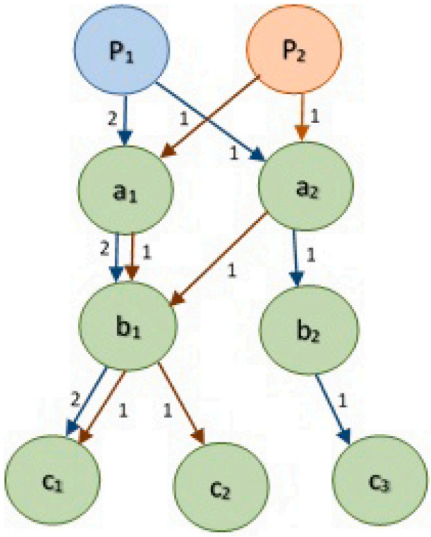


Fig. 3. Weighted MultiTree for sample provider details.

3.1.1. Weights and threshold calculation

Each unique profile information and service detail together is considered as a unique details-set and the claims of all providers having similar details-set are considered as a group of similar claims. In WMT, each complete path from a root node to the leaf node is considered as a details-set of a provider type by which the root node is formed. From Fig. 3, the path $P_1 \rightarrow a_1 \rightarrow b_1 \rightarrow c_1$ is one details-set of provider type P_1 and two providers have the same details-set. The claims of those two providers form a group of similar claims. In this work, we assign the weight to each details-set by calculating the relative frequency as the ratio of the accumulated frequency of a details-set to the sum of the accumulated frequency of all details-sets of a corresponding provider type. Let π_k be the details-set with k details, then π_{kp} will be the details-set of a provider type p. Then the accumulated frequency of each details-set π_{kp} of provider type p can be calculated as the sum of edge frequencies of all details of a details-set π_{kp} of provider type p as given in Eq. (1).

$$f(\pi_{kp}) = \sum_{i=2}^{k} f\left(\pi_{ip}\right) \tag{1}$$

Now, the weight of each details-set π_{kp} of provider type p can be calculated using the Eq. (2).

$$W(\pi_k) = \hat{p}(\pi_{kp}) = \frac{f(\pi_{kp})}{\sum_n f(\pi_{kp})} \tag{2}$$

where $\hat{p}(\pi_{kp})$ is relative frequency of details-set π_{kp} of provider type p and n is the number of details-sets yielded by the MultiTree for a provider type p.

3.1.2. Fraudulent service detection

The providers with details-set which has lower weight than some desired threshold value are considered having a mismatch in the profile and the rendered services. We calculate the threshold value as the sum of maximum and minimum weight among all details-sets of all provider types divided by the total number of claims and is given in Eq. (3). (The justification for this calculation is explained in the later section.)

$$thrd = \frac{(W(\pi_k))_{\_max} + (W(\pi_k))_{\_min}}{N} \tag{3}$$

where $(W(\pi_k))_{\_max}$, $(W(\pi_k))_{\_min}$ are the maximum and minimum weights among all calculated weights of the details-sets and N is the total number of claims. Then based on the calculated threshold value, the services rendered by the provider with a details-set whose weight is

below the threshold value are treated as fraudulent services. Algorithm 2 describes the steps for calculating weights and thresholds and identifying the fraudulent services.

Algorithm 2. Weights, threshold calculation, detecting the services

Input: MultiTree of dataset D with N claims
Output: Claims associated with fraudulent services

for each provider type p in MultiTree **do**
navigate through all the paths connected to the root node of provider type p
list all possible paths as details-set $< \pi_{1p}, \pi_{2p}, ., \pi_{kp} >$
initialize number of details-set $n = 0$
for each details-set π_{kp} of the provider type p
calculate the frequency of a details-set as

$$f(\pi_{kp}) = \sum_{i=2}^{k} f\left(\pi_{ip}\right)$$

calculate the relative frequency of a details-set π_{kp} as

$$\hat{p}(\pi_{kp}) = \frac{f(\pi_{kp})}{\sum_n f(\pi_{kp})}$$

Assign $\hat{p}(\pi_{kp})$ as weight $W(\pi_{kp})$ to the details-set π_{kp}
end
end
for all details-sets in the MultiTree
get maximum of the weights as $(W(\pi_k))_{max}$
get minimum of the weights as $(W(\pi_k))_{min}$
calculate threshold value as

$$thrd = \frac{(W(\pi_k))_{max} + (W(\pi_k))_{min}}{N}$$

end
for any details-set π_k in MultiTree
if $W(\pi_k) < thrd$
return the claims with details-set π_k as claims associated with fraudulent services
end
end

The complexity of Algorithm 2 can be analyzed as the sum of the complexities of assigning weight to each details-set, finding maximum and minimum weight among the details-sets, and finding the details-sets with weight less than threshold. Assigning weights to the details-sets needs to traverse among each path in the WMT. Finding maximum and minimum weight among the details-sets, and finding the details-sets with less than threshold need to traverse among each details-set. Let us assume a worst case scenario of N number of unique provider types having the remaining details common as the other provider types. That is each node in the tree is connected to all of the subsequent child nodes. In the described scenario, the number of paths in the MultiTree will be (N(C-1) + (C-1)*(C-1)) where C is the number of columns considered for details. Therefore, the complexity of assigning weight to each details-set will be $T(NC + C^2)$. The number of details-sets will be (N*(C-1)*(C-1)) and hence the complexity of finding maximum and minimum weight among the details-sets, and finding the details-sets with weight less than threshold will be $T(NC^2 + NC^2)$ i.e. $T(NC^2)$. Therefore, the total complexity of Algorithm 2 will be $T(NC + NC^2 + C^2)$ i.e. $T(NC^2 + C^2)$.

3.2. Applying density based clustering techniques for detecting fraud in service counts and claimed amounts

The density based clustering techniques are preferred for detecting the outliers in the healthcare claims since the centroid based clustering techniques such as k-means clustering algorithms are only effective on smaller data sets. The centroid based clustering techniques take high time complexity to detect the outliers if the data is large since they need to iterate over all of the data points. The healthcare data is scattered as non-circular points which K-means clustering algorithms are not able to cluster correctly. The density based clustering techniques can effectively identify outliers from a set of non-circular points, even in high dimensional space. It marks the points as outliers if that point is too far from the dense regions of points. Since the service counts and claimed amounts are continuous variables that need to be analyzed based on their density of values, we have chosen to apply density based clustering algorithms. The hierarchical clustering technique such as BIRCH and distance based clustering techniques also require automating the determination of number of clusters or components. Automating the determination of number of clusters is a complex problem in clustering the data points. In addition, distance based clustering techniques result in biased solutions of clustering data points even for different partition measures (Thrun, 2021). The density based clustering techniques such as DBSCAN, OPTICS, DENCLUE, and GDBSCAN have no need of the complex task of automating the number of clusters to outlier the data points. Even though the other density based clustering techniques such as Gaussian Mixing Model (GMM) is able to cluster the non-circular data points, GMM requires automating the number of clusters to be formed.

Hence, we have applied the most popular and effective density based clustering algorithms DBSCAN (Schubert, Sander, Ester, Kriegel, & Xu, 2017; Singh & Meshram, 2017) and other variants of DBSCAN including OPTICS, DENCLUE, and GDBSCAN which are non-parametric algorithms. DBSCAN clusters the dense regions of points based on the distances among them. Any distance function can be used to calculate the distance among points. DBSCAN also requires to tune two more core parameters: eps (ε) which specifies the radius of the neighborhood concerning some point and $minPts$ which is a parameter that specifies the minimum number of points that should be in ε-neighborhood of a point to consider the dense region as a cluster. However, DBSCAN requires to choose the values for the parameters eps and $minPts$ carefully since too small values will cause a large part of the data not to be clustered, whereas too high values will merge outliers and the data objects into the same cluster. The values of the eps and $minPts$ are tightly coupled with the distance function chosen for the data which can greatly impact the performance of the clustering. In this study, we have used Euclidean distance to measure the distance among the services count and the claimed amounts to identifying the nearest neighboring values. The trade-off between different parameters is performed using sensitivity analysis to decide the values that give the better performance.

The variants OPTICS, DENCLUE, and GDBSCAN also require the parameters same as DBSCAN except for the distance calculation. OPTICS algorithm uses its in-build functions to estimate the distance among the points. It has also been stated that DENCLUE is faster and GDBSCAN is efficient than DBSCAN in outlying the deviated behaviors (Singh & Meshram, 2017). The performance of each algorithm greatly varies according to the structure of the data. Hence these four algorithms have experimented with health insurance claims to compare and conclude the best suitable algorithm that can be used in the second stage. Algorithm 3 is depicting the steps of applying density based clustering algorithms on the service counts and the claimed amounts for the rendered services by a group of providers with similar profiles and service details. Algorithm 3 considers each group of similar claims i.e. all the claims related to one details-set from the WMT obtained from the first stage. Then it estimates the optimized values for parameters eps and $minPts$ based on the structure of the service counts and claimed

amounts of each group. Euclidean distance is taken as the $distFunc$ for the DBSCAN algorithm. Then it applies clustering algorithms on the service counts and the claimed amounts individually on each group of claims to detect fraudulent activities of providers.

Algorithm 3. Applying density based clustering algorithms on service counts and claimed amounts for the rendered services by a group of providers with similar profiles and service details

Input: MultiTree obtained from the first stage. **Output:** fraudulent activities detected by DBSCAN, OPTICS, DENCLUE, GDBSCAN algorithms

for each group of claims ϕ **do**
  apply density based clustering algorithms on services count
  $distFunc$ = Euclidean_distance
  $outliers\_dbscan\_sc$ = DBSCAN (ϕ, $distFunc$, $estimate\_eps$, $minPts$)
  $outliers\_optics\_sc$ = OPTICS (ϕ, eps, $minPts$)
  $outliers\_denclue\_sc$ = DENCLUE (ϕ, eps, $minPts$)
  $outliers\_gdbscan\_sc$ = GDBSCAN (ϕ, eps, $minPts$)
end
return $outliers\_dbscan\_sc$
return $outliers\_optics\_sc$
return $outliers\_denclue\_sc$
return $outliers\_gdbscan\_sc$
for each group of claims ϕ **do**
  apply density based clustering algorithms on claimed amounts
  $distFunc$ = Euclidean_distance
  $outliers\_dbscan\_ca$ = DBSCAN (ϕ, $distFunc$, $estimate\_eps$, $minPts$)
  $outliers\_optics\_ca$ = OPTICS (ϕ, eps, $minPts$)
  $outliers\_denclue\_ca$ = DENCLUE (ϕ, eps, $minPts$)
  $outliers\_gdbscan\_ca$ = GDBSCAN (ϕ, eps, $minPts$)
end
return $outliers\_dbscan\_ca$
return $outliers\_optics\_ca$
return $outliers\_denclue\_ca$
return $outliers\_gdbscan\_ca$

3.3. Description of WMT using a dataset

The proposed healthcare fraud detection model has experimented on the claims submitted by the providers registered under the CMS part B (CMS, Medicare provider utilization and payment data) program. The Centre for Medicare and Medicaid Services (CMS), is a USA based healthcare program that manages and maintains healthcare utilization by providers on behalf of subscribers for rendering various services. The CMS dataset contains around 80 features describing the details of providers, services, medicare drugs, beneficiaries, and their submitted claims. Since the proposed method is to analyze and detect provider frauds such as fraudulent services, rendering unnecessary services, and the excessive charges for rendered services, we have considered only the providers' provider information, rendered service details of the providers, and the claims submitted by the providers to develop our model. We have taken the claims submitted in the year 2018 which contains approximately 6452 distinct procedure codes categorize into 95 provider types performed by 1,12,217 providers from worldwide and has a total of 10,48,575 instances with 26 major features related to providers. CMS developed the NPPES (National Plan & Provider Enumeration System) to assign unique identifiers, known as National Provider Identifiers (NPIs), to healthcare providers and to collect the provider's profiles information such as name, credentials, gender, address, and entity type at the time of enrollment.

The CMS allows the claims from the registered providers where the details of the claims are included with the Healthcare Common Procedure Coding System (HCPCS) code, the total services count, place of service, services count per beneficiary, average submitted charge amounts. The detailed description of different features of the CMS Part B health insurance claims dataset can be known from a methodology overview published by CMS.[8] A description of the CMS Part B dataset features related to the providers' profile information, service details, and claims is given in Appendix A. In our model, the provider profiles such as provider type, credential of the provider, entity type, zip code, and the service details such as place of service, HCPCS code are considered in the first stage for analyzing the similarity among providers as these features can uniquely identify each claim. The sample CMS data with the considered features are shown in Table 1.

In the first stage, we have constructed WMT for claims data based on the details of the providers and services to group the providers with similar details as depicted in Algorithm 1. Fig. 4 is a sample WMT for CMS Part B claims with the details of the providers whose provider types are "pathology" and "physician medicine and rehabilitation". When constructing the WMT, all the claims are entered in the same order of details in which the first claim details are entered. For example, in the WMT of Fig. 4, the order of details is provider type, credentials, entity type, place of service, Zip code, HCPCS code. The zeroth level of the WMT i.e the root nodes are the provider types "pathology" and "physician medicine and rehabilitation". The nodes of the next level of the tree are the credentials of the providers. All the credentials associated with "pathology" and "physician medicine and rehabilitation" in the claims data are considered as the nodes in the first level connected to the associated provider type node. If there are common credentials for both provider types such as D.O., D.O.FSAP, M.D., in the sample WMT, those nodes will be shared by the two provider types instead of having redundant nodes in the first level for the same credentials. In such a way, all the claims can be constructed as WMT for the details defined by the users. The frequency of occurrences is assigned to each edge between the details based on their occurrence in the claims data. Every single path from the root node to the leaf node is one details-set for the corresponding provider type. For example, from Fig. 4. Pathology → D.O.FSAP → I → F → Y → J1040 is one details-set for provider type "pathology".

The weight of each details-set yielded from the WMT and the threshold value can be calculated as depicted in Algorithm 2. If the weight of the service associated with the specific provider details and service details is less than the threshold value, that service will be defined as a fraudulent service. Then the second stage of our model analyzed the claims of providers' which is of continuous type. The number of services provided per day, and the submitted charges per service are the major claims in CMS part B data that describe the deviations if any in the behavior of providers. Hence, the second stage of our model analyzed the service counts and claimed amounts of the providers with similar profiles and services. Due to the space limitation, we are presenting three clusters each for the service counts and the submitted charges of the providers with similar details. Consider that ds_1, d_2, and ds_3 are the three different non-fraudulent details-sets of the providers where ds_1 has details as provider type "pathology", Credentials "M.D.", Entity Type "I", Place of Service "F", Zip Code "370 277 541", and HCPCS code "88 313"; ds_2 has provider type "pathology", Credentials "M.D.", Entity Type "I", Place of Service "F", Zip Code "212 870 005", and HCPCS code "88 341"; and ds_3 has provider type "Diagnostic Radiology", Credentials "D.O.", Entity Type "I", Place of Service "F", Zip Code "432 143 937", and HCPCS code "71 010". Then the matrix in Fig. 5 represents the cluster of the services count and the submitted charges for three different details-sets ds_1, ds_2, ds_3.

The number of claims in each group of similar claims ranges from 50 to 2000 based on the profiles of the providers. This indicates that the providers with some provider types have less number of cases

[8] https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Medicare-Physician-and-Other-Supplier-PUF-Methodology.pdf.

Table 1
Sample CMS data.

NPI	NPPES credentials	Provider type	HCPCS code	Place of service	Number of services	Average medicare allowed amount
1063403947	MD	Internal medicine	99214	O	155	120.15
1063403947	MD	Internal medicine	99231	F	21	41.68
1063403947	MD	Internal medicine	99232	F	13	76.73
1063404085	MD	Pathology	88305	O	776	44.25595
1063404085	MD	Pathology	88307	F	139	78.23

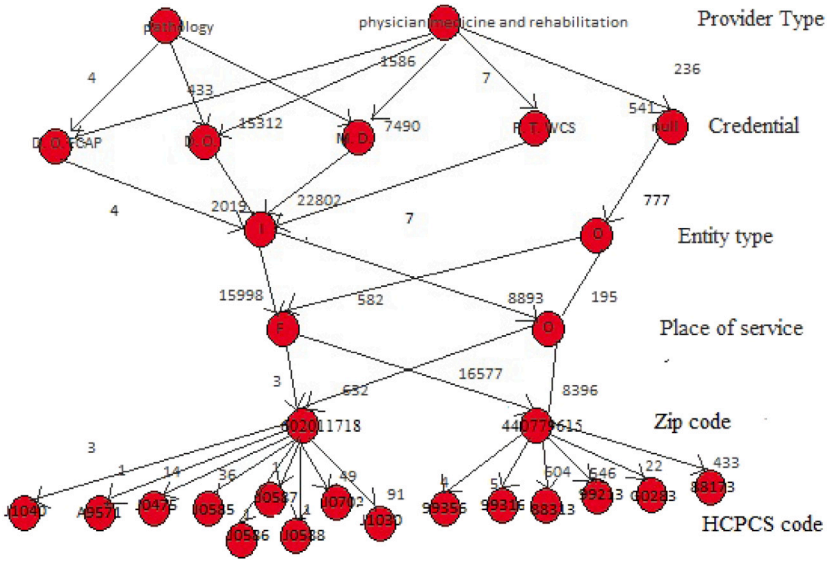


Fig. 4. Weighted MultiTree with sample details from CMS part B 2018.

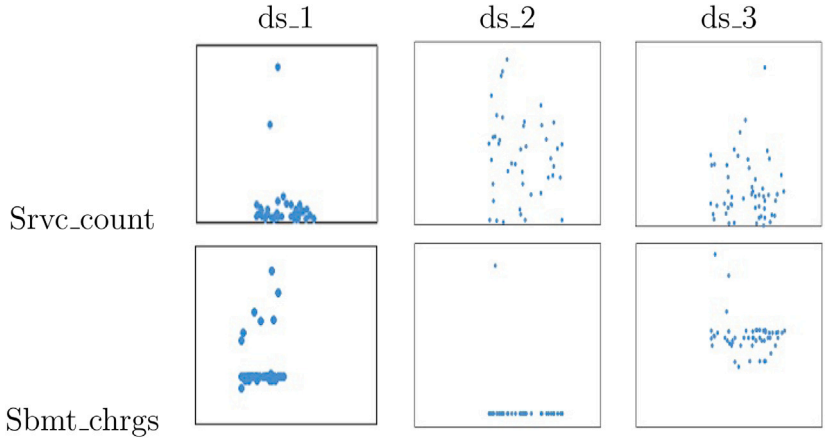


Fig. 5. Sample groups of claims for providers with similar details.

and the providers with some provider types like "general medicine" provide a huge number of services per annum. Once the group of claims with similar provider details is formed, four density based clustering algorithms DBSCAN, OPTICS, DENCLUE, and GDBSCAN are applied to detect the fraudulent activities of providers from each group of claims. The fraudulent activities of providers can be defined as the claims that have a deviation in amounts or counts among the providers with similar details. However, there might be redundancy in fraudulent claims detected in the first stage and by each clustering algorithm in the second stage. Hence the fraudulent claims from two stages without redundancy are considered as the total number of fraudulent claims detected using our proposed model. The accuracy of our model is measured by comparing the fraudulent claims detected with real-world fraudulent cases.

As the CMS part B dataset is not labeled with the fraudulent cases, the real-world fraudulent cases can be known from the LEIE (List of Exclusions of Individual Entities) dataset, which have been updated every month by the Office of Inspection General (OIG). LEIE also includes amendments as exclusion types that categorize the fraudulent cases into the associated type of fraud and also aids to know the cause for denying the claims for reimbursement. The sample LEIE dataset and some of the amendments or exclusion types are presented in Table 2 and Fig. 6 respectively. From the list of mentioned amendments, we have considered the excluded claims of the LEIE dataset which have social security act number 1128(b)(6) to compare with the fraudulent services detected using our model. We found 2156 instances as the number of fraudulent claims through comparing excluded claims of LEIE with CMS claims.

Social Security Act	42 USC §	Amendment
1128(b)(1)(A)	1320a-7(b)(1)(A)	Misdemeanor conviction relating to health care fraud. Baseline Period: 3 years
1128(b)(1)(B)	1320a-7(b)(1)(B)	Conviction relating to fraud in non-health care programs. Baseline Period: 3 years
1128(b)(2)	1320a-7(b)(2)	Conviction relating to obstruction of an investigation or audit. Baseline Period: 3 years
1128(b)(3)	1320a-7(b)(3)	Misdemeanor conviction relating to controlled substance. Baseline Period: 3 years
1128(b)(4)	1320a-7(b)(4)	License revocation, suspension, or surrender. Minimum Period: Period imposed by the state licensing authority.
1128(b)(5)	1320a-7(b)(5)	Exclusion or suspension under federal or state health care program. Minimum Period: No less than the period imposed by federal or state health care program.
1128(b)(6)	1320a-7(b)(6)	Claims for excessive charges, unnecessary services or services which fail to meet professionally recognized standards of health care, or failure of an HMO to furnish medically necessary services. Minimum Period: 1 year

Fig. 6. Social security act numbers and amendment details of some fraud types..

Table 2
Sample of February 2019 LEIE data.

Specialty	NPI	City	EXCLTYPE	EXCLDATE
Podiatry practice	1598041998	Foresthills	1128a1	20190320
Pharmacy	1275750374	Lynbrook	1128a1	20190320
Transportation Co	0	Phoenix	1128a1	20190320
Adult Day Care Facil	0	Santa Rosa	1128a1	20190320

3.4. Performance metrics

The metrics used in our model for evaluating the performance are False Alarm Ratio (FAR) which can be defined as the ratio of the number of claims grouped incorrectly to the total number of claims grouped by the model under one provider type, Fraud Detection Accuracy (FDA) which can be measured as the ratio of the total number of fraudulent claims detected correctly by the model to the total number of actual fraudulent claims, Precision, Recall, F-score, and overall accuracy. The complete description for the performance metrics is given in Appendix B.

4. Results and analysis

In this section, we present the results to explain the performance improvement in detecting fraudulent claims using our proposed model. In particular, we present the analysis to evidence our assertion that the WMT approach enhances the performance in analyzing similar claims. This section is organized according to the flow of our model which presents the results in the order of similarity analysis among the claims and then the fraudulent claims detection.

4.1. Multivariate analysis for grouping similar claims

We analyzed the similarity among the claims based on the details-sets identified using WMT where each details-set contains the unique information of profile and geospatial location details and services of providers. We consider that the claims belong to a details-set are similar claims and hence we measure the performance of similarity analysis in terms of the number of details-set exactly detected by WMT for each provider type. Table 3 has presented the actual number of unique details-sets (including fraudulent and non-fraudulent) under each provider type and the number of details-sets correctly detected as fraudulent and non-fraudulent by WMT. The actual number of unique details-sets of each provider type are measured based on the claims of the CMS part B dataset. And We present the number of details-sets detected for the top 25 provider types from CMS claims data.

Our model has detected the details-sets as fraudulent if there is any mismatch in the provider profile and the rendered services that analyzed based on the frequency of occurrences. Whether a details-set is fraudulent or non-fraudulent, the total number of unique details-sets detected by our approach is almost similar to the number of actual instances. The wrong predictions in grouping the similar claim might have occurred due to the existence of very few claims associated with the details set. The improvement in the similarity analysis using our approach is evidenced by performing the comparative analysis of our similarity analysis results with the other approaches Mahalabonis Distance (DB), Density-Based Analysis using k Nearest Neighbors (DBA-kNN), and Neural Networks with backpropagation (NNbp) which were commonly used for similarity analysis in existing fraud detection models. The multivariate data which was used to analyze the similarity among the claims using our approach WMT is taken as the input features for DB, DBA-kNN, and NNbp for ease in comparison. We

Table 3
Results of WMT in grouping similar claims.

Provider type	Actual # unique detail-sets	Correctly detected	
		Non-fraudulent	Fraudulent
Diagnostic radiology	2065	2037	15
Internal medicine	1830	1818	11
Family practice	1783	1781	2
Nurse practitioner	1179	1173	3
Cardiology	757	756	0
Physician assistant	762	759	3
Orthopedic surgery	556	551	2
Physical therapist	498	495	2
Anesthesiology	432	432	0
Ophthalmology	448	446	1
Emergency medicine	361	358	1
Dermatology	403	401	0
Podiatry	400	396	0
Urology	351	350	1
Gastroenterology	331	329	1
Optometry	318	316	0
Hematology-Oncology	310	309	1
General surgery	297	297	0
Pathology	276	274	1
Pulmonary disease	268	261	3
Neurology	248	244	2
CRNA	188	187	1
Clinical laboratory	181	176	4
Nephrology	213	212	0
Obstetrics & Gynecology	181	180	1

Table 4
Results of WMT in grouping similar claims.

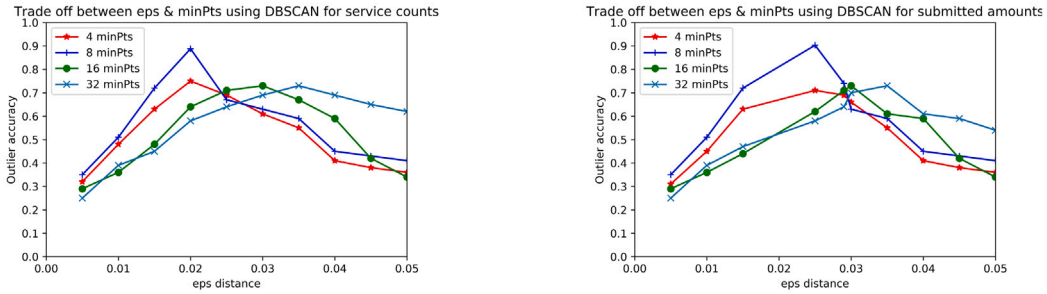
Provider type	MD	DBA-kNN	NNbp	WMT
Diagnostic radiology	0.028	0.036	0.078	0.017
Internal medicine	0.057	0.071	0.028	0.007
Family practice	0.028	0.064	0.077	0.024
Nurse practitioner	0.065	0.077	0.043	0.002
Cardiology	0.028	0.034	0.074	0.021
Physician assistant	0.074	0.049	0.077	0.011
Orthopedic surgery	0.031	0.041	0.044	0.007
Physical therapist	0.081	0.071	0.079	0.005
Anesthesiology	0.067	0.083	0.051	0.009
Ophthalmology	0.083	0.033	0.027	0.003
Emergency medicine	0.053	0.037	0.027	0.008
Dermatology	0.081	0.073	0.057	0.017
Podiatry	0.053	0.068	0.038	0.007
Urology	0.075	0.052	0.056	0.014
Gastroenterology	0.044	0.041	0.054	0.014
Optometry	0.070	0.053	0.073	0.003
Hematology-Oncology	0.051	0.033	0.074	0.004
General surgery	0.039	0.068	0.085	0.012
Pathology	0.072	0.031	0.071	0.018
Pulmonary disease	0.053	0.042	0.068	0.011
Neurology	0.061	0.029	0.068	0.021
CRNA	0.054	0.074	0.068	0.020
Clinical laboratory	0.081	0.083	0.074	0.006
Nephrology	0.034	0.064	0.028	0.001
Obstetrics & Gynecology	0.032	0.062	0.058	0.017

measure the performance of models per each provider type in terms of False Alarm Ratio (FAR). The less value of FAR indicates the better performance of the model in grouping similar claims. According to that, the results of FAR presented in Table 4 depicting that our WMT approach has shown a significant improvement in grouping similar claims.

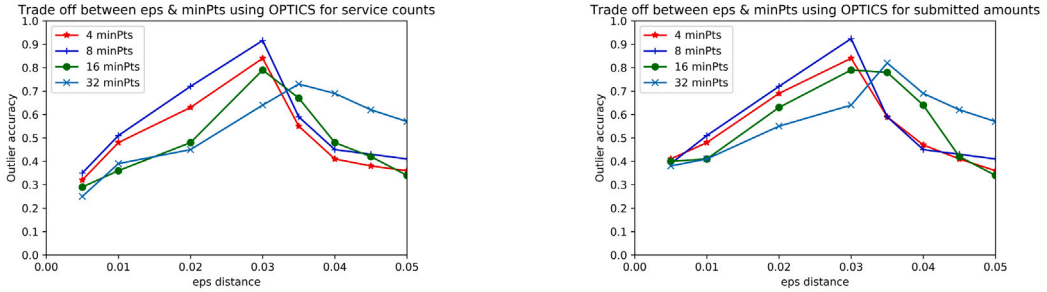
4.2. Fraudulent claims detection

The fraudulent claims in the second stage can be detected through density based clustering which requires parameterization of eps and $minPts$. The choice of these parameters is greatly dependent on the structure of the data and the optimized values the performance of the detection system can be enhanced. Hence, we have conducted a trade-off analysis between eps and $minPts$ in order to get better accuracy for proposed density based algorithms. The trade-off analysis for the parameters eps and $minPts$ for CMS part B claims are presented in Fig. 7. As depicted in Fig. 7(a), the overall accuracy is better for eps = 0.02, $minPts$ = 8 & eps = 0.026, $minPts$ = 8 when DBSCAN is applied on the service counts and submitted charges respectively. The trade-off analysis results from Fig. 7(b) concludes that the overall accuracy is better for eps = 0.029, $minPts$ = 8 & eps = 0.031, $minPts$ = 8 when OPTICS is applied on the service counts and submitted charges respectively. The results from Fig. 7(c) concludes that the overall accuracy is better for eps = 0.03, $minPts$ = 4 & eps = 0.037, $minPts$ = 4 when DENCLUE is applied on the service counts and submitted charges respectively and the results from Fig. 7(d) concludes that the overall accuracy is better for eps = 0.032, $minPts$ = 8 & eps = 0.035, $minPts$ = 8 when GDBSCAN is applied on the service counts and the submitted charges respectively.

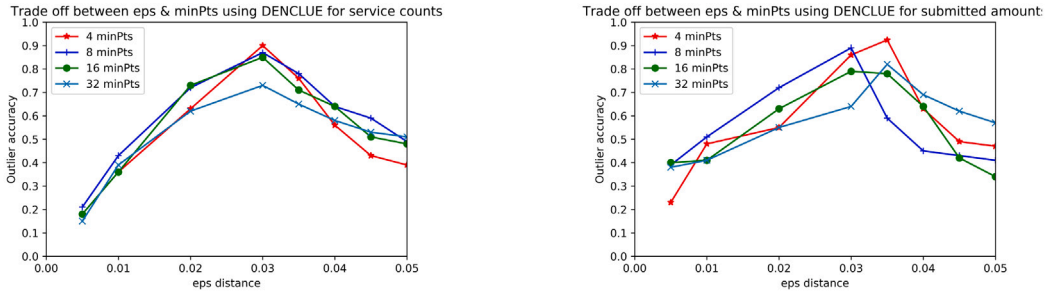
Density based clustering algorithms with the derived parameter values are applied on each group of claims of CMS part B data to analyze the fraud detection performance of our model. The performance is measured in terms of Fraud Detection Accuracy (FDA) for four approaches WMT with DBSCAN (WMT-DBSCAN), WMT with OPTICS (WMT-OPTICS), WMT with DENCLUE (WMT-DENCLUE), and WMT with GDBSCAN (WMT-GDBSCAN) are compared with the existing healthcare fraud detection models LR-COF, LR-SOF, MNB-COF, and MNB-SOF developed by Herland et al. (2018). We have chosen these models for performance analysis since these models are also developed similar to our experiment flow which analyzed the similarity among the claims based on the provider types and then applied ML techniques for classifying fraudulent and non-fraudulent claims. We have applied these existing models on service counts and submitted charges and measured the FDA values for each provider type. Random undersampling is applied on service counts and submitted charges of each group as the healthcare claims data is imbalanced with very few fraudulent cases compared to the non-fraudulent. To avoid the biased results due to undersampling, FDA values are recorded for 20 runs of 10-folds. The average value of 20 runs of 10-fold cross validation is considered as the FDA value for each group of claims. We compared the obtained FDA results of our model with the existing healthcare fraud detection



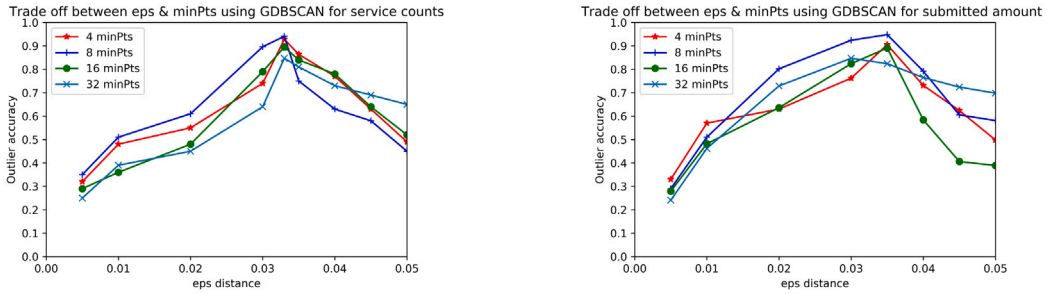
(a) Parameterization for DBSCAN based on the claims of service counts and submitted charges.



(b) Parameterization for OPTICS based on the claims of service counts and submitted charges.



(c) Parameterization for DENCLUE based on the claims of service counts and submitted charges.



(d) Parameterization for GDBSCAN based on the claims of service counts and submitted charges.

Fig. 7. Tradeoff between the parameters *eps* and *minPts* for density based clustering algorithms.

models. The comparative analysis for service counts and submitted charges is presented in Tables 5 and 6 respectively.

The comparative results of different fraud detection models are shown in Tables 5 and 6 for service counts and submitted charges depicting that our approaches WMT-DBSCAN, WMT-OPTICS, WMT-DENCLUE, and WMT-GDBSCAN have some improvement in detection accuracy. The existing approach MNB-SOF showed some better results for a few provider types as it was developed based on supervised learning. However, our proposed approaches showed significant improvement in the results for maximum groups associated with different provider types. Among the four proposed approaches, WMT-GDBSCAN exhibited better performance as per the FDA results.

We also evaluated our model using the metrics precision, recall, F-score, and overall accuracy in order to analyze the performance of our proposed approaches in terms of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The definitions for TP, TN, FP, and FN in our context are given as below:

TP: Fraudulent claims detected as fraudulent using the model.

TN: Non-Fraudulent claims detected as non-fraudulent claims using the model.

Table 5
Comparative analysis for the performance in detecting fraudulent claims (service counts with random undersampling).

Provider type	LR-COF	LR-SOF	MNB-COF	MNB-SOF	WMT-DBSCAN	WMT-OPTICS	WMT-DENCLUE	WMT-GDBSCAN
Diagnostic radiology	0.894	0.823	0.868	0.876	0.912	0.853	0.923	0.813
Internal medicine	0.843	0.887	0.905	0.867	0.846	0.823	0.833	0.916
Family practice	0.823	0.811	0.834	0.894	0.942	0.846	0.814	0.825
Nurse practitioner	0.817	0.822	0.866	0.864	0.889	0.884	0.820	0.951
Cardiology	0.818	0.907	0.801	0.833	0.954	0.843	0.894	0.955
Physician assistant	0.842	0.879	0.823	0.878	0.907	0.821	0.958	0.900
Orthopedic surgery	0.899	0.885	0.863	0.839	0.874	0.978	0.881	0.919
Physical therapist	0.894	0.809	0.906	0.911	0.833	0.911	0.877	0.943
Anesthesiology	0.832	0.807	0.811	0.812	0.834	0.932	0.854	0.932
Ophthalmology	0.870	0.851	0.896	0.869	0.944	0.901	0.868	0.881
Emergency medicine	0.819	0.803	0.846	0.825	0.854	0.923	0.847	0.863
Dermatology	0.804	0.837	0.845	0.837	0.859	0.943	0.913	0.945
Podiatry	0.853	0.884	0.902	0.861	0.843	0.872	0.843	0.859
Urology	0.893	0.811	0.857	0.825	0.953	0.917	0.904	0.948
Gastroenterology	0.823	0.817	0.876	0.809	0.826	0.901	0.869	0.935
Optometry	0.883	0.863	0.842	0.826	0.901	0.864	0.879	0.953
Hematology-Oncology	0.883	0.854	0.824	0.801	0.903	0.926	0.812	0.842
General surgery	0.845	0.842	0.937	0.863	0.846	0.899	0.913	0.955
Pathology	0.817	0.823	0.826	0.909	0.843	0.845	0.864	0.917
Pulmonary disease	0.836	0.843	0.815	0.859	0.857	0.900	0.899	0.849
Neurology	0.877	0.804	0.839	0.857	0.853	0.842	0.903	0.945
CRNA	0.846	0.826	0.801	0.874	0.946	0.836	0.848	0.912
Clinical laboratory	0.811	0.815	0.889	0.863	0.926	0.913	0.937	0.954
Nephrology	0.901	0.864	0.801	0.834	0.873	0.833	0.952	0.892
Obstetrics & Gynecology	0.854	0.843	0.853	0.854	0.832	0.853	0.902	0.857

Table 6
Comparative analysis for the performance in detecting fraudulent claims (submitted amounts with random undersampling).

Provider type	LR-COF	LR-SOF	MNB-COF	MNB-SOF	WMT-DBSCAN	WMT-OPTICS	WMT-DENCLUE	WMT-GDBSCAN
Diagnostic radiology	0.902	0.919	0.825	0.829	0.947	0.954	0.946	0.918
Internal medicine	0.902	0.847	0.831	0.833	0.918	0.853	0.918	0.927
Family practice	0.857	0.802	0.846	0.837	0.863	0.822	0.917	0.862
Nurse practitioner	0.851	0.842	0.847	0.811	0.926	0.852	0.899	0.947
Cardiology	0.811	0.879	0.892	0.817	0.857	0.919	0.926	0.948
Physician assistant	0.879	0.913	0.836	0.827	0.817	0.817	0.918	0.871
Orthopedic surgery	0.903	0.865	0.873	0.917	0.865	0.937	0.833	0.945
Physical therapist	0.876	0.881	0.876	0.902	0.856	0.852	0.843	0.865
Anesthesiology	0.893	0.804	0.807	0.835	0.911	0.923	0.897	0.946
Ophthalmology	0.876	0.812	0.816	0.873	0.911	0.915	0.903	0.959
Emergency medicine	0.802	0.863	0.816	0.917	0.923	0.917	0.918	0.836
Dermatology	0.809	0.818	0.872	0.867	0.827	0.936	0.928	0.903
Podiatry	0.907	0.867	0.800	0.824	0.874	0.924	0.847	0.836
Urology	0.822	0.869	0.852	0.887	0.883	0.897	0.846	0.907
Gastroenterology	0.857	0.887	0.798	0.918	0.875	0.869	0.873	0.845
Optometry	0.827	0.856	0.894	0.835	0.927	0.947	0.817	0.927
Hematology-Oncology	0.846	0.834	0.822	0.802	0.827	0.893	0.827	0.947
General surgery	0.891	0.876	0.901	0.893	0.921	0.920	0.897	0.953
Pathology	0.819	0.843	0.815	0.826	0.893	0.817	0.937	0.926
Pulmonary disease	0.816	0.897	0.847	0.910	0.863	0.875	0.911	0.874
Neurology	0.903	0.853	0.837	0.859	0.928	0.888	0.917	0.929
CRNA	0.883	0.827	0.847	0.803	0.894	0.826	0.827	0.917
Clinical laboratory	0.910	0.902	0.856	0.865	0.876	0.923	0.876	0.821
Nephrology	0.801	0.823	0.914	0.803	0.889	0.879	0.926	0.904
Obstetrics & Gynecology	0.854	0.897	0.846	0.899	0.875	0.913	0.856	0.943

FP: Fraudulent claims detected as non-fraudulent using the model.

FN: Non-fraudulent claims detected as fraudulent using the model.

We combined the fraudulent claims detected in each group by the existing approaches LR-COF, LR-SOF, MNB-CF, MNB-SOF, and our four approaches WMT-DBSCAN, WMT-OPTICS, WMT-DENCLUE, and WMT-GDBSCAN for measuring the metrics by mapping the fraudulent claims detected by the model with real-world fraudulent claims to measure the metrics. We also compared the results with other existing models MARS-BI, GAN-LR, GAN-XGB, GAN-RF, and GAN-DT which analyzed the fraudulent claims without any similarity analysis among the claims. The analysis results for detecting fraudulent claims in service counts and submitted charges are presented in Tables 7 and 8 respectively. The analysis is evidencing that the models which analyze fraudulent claims by grouping similar claims have significant improvement in fraud detection accuracy. From the analysis of the results, it is observed that the proposed approaches enhanced the performance in detecting fraudulent service counts and submitted charges submitted by the providers. We observed the high overall accuracies of 0.941, 0.938, 0.920, and 0.944 for detecting fraudulent service counts and 0.939, 0.942, 0.937, and 0.946 for detecting fraudulent submitted charges using our approaches WMT-DBSCAN, WMT-OPTICS, WMT-DENCLUE, and WMT-GDBSCAN respectively. We also observed that among the proposed approaches, WMT-GDBSCAN outperformed the remaining proposed approaches and the existing models.

5. Conclusion

Healthcare fraud can be defined as the fraudulent behaviors of some health providers or organizations degrading the quality of health services. In this paper, we have proposed an unsupervised multivariate analysis model to detect fraudulent behaviors in health insurance claims. Our model consisted of two stages in which the first stage

Table 7
Performance analysis for fraudulent claims detection in service counts using overall accuracy.

Method	Precision	Recall	F-score	Overall accuracy
LR-COF	0.864	0.852	0.857	0.873
LR-SOF	0.927	0.875	0.900	0.917
MNB-COF	0.864	0.851	0.857	0.891
MNB-SOF	0.907	0.865	0.885	0.922
MARS-BI	0.889	0.873	0.880	0.861
GAN-LR	0.841	0.853	0.846	0.851
GAN-XGB	0.811	0.826	0.818	0.869
GAN-RF	0.853	0.853	0.853	0.847
GAN-DT	0.832	0.837	0.834	0.903
WMT-DBSCAN	0.923	0.911	0.916	0.941
WMT-OPTICS	0.936	0.922	0.928	0.938
WMT-DENCLUE	0.921	0.893	0.906	0.920
WMT-GDBSCAN	0.957	0.958	0.957	0.944

Table 8
Performance analysis for fraudulent claims detection in service counts using overall accuracy.

Method	Precision	Recall	F-score	Overall accuracy
LR-COF	0.876	0.813	0.843	0.867
LR-SOF	0.899	0.879	0.888	0.913
MNB-COF	0.824	0.863	0.843	0.898
MNB-SOF	0.932	0.886	0.908	0.923
MARS-BI	0.818	0.878	0.846	0.901
GAN-LR	0.877	0.894	0.885	0.895
GAN-XGB	0.843	0.822	0.32	0.908
GAN-RF	0.876	0.829	0.851	0.861
GAN-DT	0.917	0.871	0.893	0.857
WMT-DBSCAN	0.953	0.932	0.942	0.939
WMT-OPTICS	0.932	0.937	0.934	0.942
WMT-DENCLUE	0.912	0.889	0.900	0.937
WMT-GDBSCAN	0.945	0.937	0.940	0.946

constructed a Weighted MultiTree to group the similar claims and detect the fraudulent services. The second stage applied popular density based clustering algorithms such as DBSCAN, OPTICS, DENCLUE, and GDBSCAN on each group of claims of similar providers to detect the unnecessary services and the excessive charges submitted by the providers for the rendered services. The performance of our model is evaluated and the obtained results illustrated that the WMT with GDBSCAN clustering outperformed in detecting fraudulent claims. The limitation of the proposed model is misjudging the details-set with a low frequency of claims transactions as fraudulent. As a future direction of this work, the problem of detection fraud for low frequency transactions by reviewing the similar works as developed in Zhang, Chen, Liu, and Wang (2020). The scope of this can be extended to develop adaptive fraud detection models based on drift analysis (Priya & Uthra, 2021; Sahmoud & Topcuoglu, 2020) and secure AI-driven models using cloud technologies (Shanmugapriya & Kavitha, 2019) and blockchain techniques (Hasselgren, Kralevska, Gligoroski, Pedersen, & Faxvaag, 2020; McGhin, Choo, Liu, & He, 2019).

CRediT authorship contribution statement

Lavanya Settipalli: Conceptualization, Methodology, Software, Investigation, Writing – original draft, Writing – review & editing, Visualization. **G.R. Gangadharan:** Conceptualization, Methodology, Writing – review & editing, Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data used for this paper is available in public domain.

Acknowledgments

This work is partially supported by the Scheme for Promotion of Academic and Research Collaboration (SPARC), sponsored by the Ministry of Human Resource Development, Government of India, under the project titled Digital Health Records Storage and Analysis for Healthcare Provisioning of Global Patients: An India - Australia Initiative (1406).

Code availability (software application or custom code):

Custom code developed.

Appendix A. Description of CMS Part B dataset features.

See Table A.9.

Appendix B. Performance metrics

The performance metrics used in this study are aided to well-explain how the proposed model can improve the accuracy in detecting fraudulent claims. The metrics used for evaluating our model are False Alarm Ratio (FAR), False Detection Accuracy (FDA), precision, recall, F-score, and overall accuracy.

B.1. False Alarm Ratio (FAR)

In general, the FAR is defined as the number of false positives. In our study, we have used the FAR to evaluate the similarity analysis of models to express the claims classified into a wrong group. We formulated the FAR for our evaluation as given in Eq. (B.1).

$$FAR = \frac{\sum_{i=1}^{K} |\varphi_{ip}| - \sum_{i=1}^{K} |\varphi_{ipc}|}{\sum_{i=1}^{K} |\varphi_{ip}|} \tag{B.1}$$

where φ_{ip} is a set of claims associated with provider type p, classified into group i by the model and $|\varphi_{ip}|$ is the total number of claims in the set φ_{ip}. φ_{ipc} is a set of correctly classified claims into group i associated with provider type p and $|\varphi_{ipc}|$ is the total number of claims classified by the model into the correct group. And K is the total number of groups with the provider type p.

B.2. Fraud Detection Accuracy (FDA)

As our model stressing about detecting fraudulent claims, we have evaluated our model in terms of the number of frauds detected by our model accurately. Hence, we formulated the value for FDA as given in Eq. (B.2).

$$FDA = \frac{N_{pcf}}{N_{af}} \tag{B.2}$$

where N_{pcf} is the number of fraudulent claims correctly detected out of the total number of fraudulent claims N_{pf} detected by the proposed model.

Table A.9
Description of CMS Part B dataset features.

Feature	Description	Datatype
npi	10 digit unique identification number of provider	String
nppes_entity_ code	Determines whether a provider is individual or organization	Categorical
nppes_provider_ last_org_name	Last name of the provider if the entity type is individual, otherwise it is organization name.	String
nppes_provider_ first_name	First name of the provider if the entity type is individual, otherwise it is empty	String
nppes_provider_ mi	Middle initial of the provider if the entity type is individual otherwise it is empty	String
nppes_credentials	Describes the specialty of the provider if the entity type is individual otherwise it is empty	Categorical
nppes_provider_ gender	Gender of the provider if the entity type is individual otherwise it is empty	Categorical
nppes_provider_ street1	Describes the street details in address of the provider.	Categorical
nppes_provider_ street2	Describes the street details in address of the provider.	Categorical
nppes_provider_ city	Describes the city details in address of the provider.	Categorical
nppes_provider_ zip	Describes the Zipcode details in address of the provider.	Categorical
nppes_provider_ state	Describes the state details in address of the provider.	Categorical
nppes_provider_ country	Describes the country details in address of the provider.	Categorical
provider_type	Derived based on the specialty of the provider that reported the claim. If two or more specialties are mentioned by the provider, it will be associated with the largest number of claims.	Categorical
medicare_ participation_ indicator	Identifies whether the provider participates in Medicare and/or accepts assignment of Medicare allowed amounts.	Categorical
place_of_service	Identifies whether the place of service submitted on the claims is a facility or non-facility	Categorical
hcpcs_code	5 digit code that identifies the specific medical service furnished by the provider	Categorical
hcpcs_description	Description of the associated HCPCS code furnished by the provider	Categorical
line_srvc_cnt	This count indicates the total number of services associated with one HCPCS code provided by the provider. This count is associated with HCPCS code and varies from service to service.	Discrete
bene_unique_cnt	Number of distinct beneficiaries received the service.	Discrete
bene_day_srvc_cnt	A beneficiary may receive the same service multiple times in a single day. This is the count that removes the duplicates and presents the unique number of services that occurred in a single day.	Discrete
average_ Medicare_ allowed_amt	Average medicare amount allowed for an HCPCS service and place of service	Categorical
average_ submitted_ chrg_amt	The amount claimed by the provider for a service	Continuous
average_ Medicare_ payment_amt	Average amount that can be paid after applying reduction percentage on the submitted amounts.	Continuous
average_ Medicare_ standardized_amt	Average amount that can be paid after standardization. Standardization removes geographic differences in payment rates such as local wages or input prices to make payments across geographic areas comparable.	Continuous

B.3. Precision, Recall, and F-score

Precision, Recall, and F-score are the common metrics that usually measure the performance of the models for all possible cases True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The formulae to measure the Precision, Recall, and F-score values are given in Eqs. (B.3), (B.4), and (B.5) respectively.

$$Precision = \frac{TP}{(TP + FP)} \tag{B.3}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{B.4}$$

$$F - score = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \tag{B.5}$$

B.4. Overall accuracy

The aforementioned FDA has calculated the accuracy of the model based on the fraudulent claims (positive case) for each group associated with different provider types. However, the overall accuracy measured the accuracy value based on two cases (positive and negative). The measurement of overall accuracy considered the fraudulent claims from entire claims data which can be obtained by combining the fraudulent claims detected for each group of claims. The formula for calculating overall accuracy is given in Eq. (B.6).

$$Overallaccuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{B.6}$$

References

Ahmad, P., Qamar, S., & Rizvi, S. Q. A. (2015). Techniques of data mining in healthcare: a review. *International Journal of Computer Applications*, *120*(15).

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod Record*, *28*(2), 49–60.

Ashtiani, M. N., & Raahemi, B. (2021). Intelligent fraud detection in financial statements using machine learning and data mining: A systematic literature review. *IEEE Access*.

Bauder, R. A., & Khoshgoftaar, T. M. (2016). A probabilistic programming approach for outlier detection in healthcare claims. In *2016 15th IEEE international conference on machine learning and applications (ICMLA)* (pp. 347–354). IEEE.

Bauder, R. A., & Khoshgoftaar, T. M. (2017). Multivariate outlier detection in medicare claims payments applying probabilistic programming methods. *Health Services and Outcomes Research Methodology*, *17*(3), 256–289.

Bauder, R., Khoshgoftaar, T. M., & Seliya, N. (2017). A survey on the state of healthcare upcoding fraud analysis and detection. *Health Services and Outcomes Research Methodology*, *17*(1), 31–55.

Bayerstadler, A., van Dijk, L., & Winter, F. (2016). Bayesian multinomial latent variable modeling for fraud and abuse detection in health insurance. *Insurance: Mathematics & Economics*, *71*, 244–252.

Boutaher, N., Elomri, A., Abghour, N., Moussaid, K., & Rida, M. (2020). A review of credit card fraud detection using machine learning techniques. In *2020 5th international conference on cloud computing and artificial intelligence: technologies and applications (CloudTech)* (pp. 1–5). IEEE.

Carrasco, R. S. M., & Sicilia-Urbán, M.-Á. (2020). Evaluation of deep neural networks for reduction of credit card fraud alerts. *IEEE Access*, *8*, 186421–186432.

Chandola, V., Sukumar, S. R., & Schryver, J. C. (2013). Knowledge discovery from massive healthcare claims data. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1312–1320).

Chelladurai, U., & Pandian, S. (2021). A novel blockchain based electronic health record automation system for healthcare. *Journal of Ambient Intelligence and Humanized Computing*, 1–11.

Dhieb, N., Ghazzai, H., Besbes, H., & Massoud, Y. (2020). A secure ai-driven architecture for automated insurance systems: Fraud detection and risk measurement. *IEEE Access*, *8*, 58546–58558.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd, Vol. 96* (pp. 226–231).

Flynn, K. (2015). Financial fraud in the private health insurance sector in Australia: Perspectives from the industry. *Journal of Financial Crime*.

Hancock, J. T., & Khoshgoftaar, T. M. (2021). Gradient boosted decision tree algorithms for medicare fraud detection. *SN Computer Science*, *2*(4), 1–12.

Haque, M. E., & Tozal, M. E. (2021). Identifying health insurance claim frauds using mixture of clinical concepts. *IEEE Transactions on Services Computing*.

Hasselgren, A., Kralevska, K., Gligoroski, D., Pedersen, S. A., & Faxvaag, A. (2020). Blockchain in healthcare and health sciences—A scoping review. *International Journal of Medical Informatics*, *134*, Article 104040.

He, H., Graco, W., & Yao, X. (1998). Application of genetic algorithm and k-nearest neighbour method in medical fraud detection. In *Asia-Pacific conference on simulated evolution and learning* (pp. 74–81). Springer.

He, H., Hawkins, S., Graco, W., & Yao, X. (2000). Application of genetic algorithm and K-nearest neighbour method in real world medical fraud detection problem. *Journal of the Advance Computer Intelligence and Intelligent Informatics*, *4*(2), 130–137.

He, H., Wang, J., Graco, W., & Hawkins, S. (1997). Application of neural networks to detection of medical fraud. *Expert Systems with Applications*, *13*(4), 329–336.

Herland, M., Bauder, R. A., & Khoshgoftaar, T. M. (2020). Approaches for identifying US medicare fraud in provider claims data. *Health Care Management Science*, *23*(1), 2–19.

Herland, M., Khoshgoftaar, T. M., & Bauder, R. A. (2018). Big data fraud detection using multiple medicare data sources. *Journal of Big Data*, *5*(1), 1–21.

Hinneburg, A., Keim, D. A., et al. (1998). An efficient approach to clustering in large multimedia databases with noise. In *Kdd, Vol. 98* (pp. 58–65).

Jiang, Z., Chen, X., Dong, B., Zhang, J., Gong, J., Yan, H., et al. (2019). Trajectory-based community detection. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *67*(6), 1139–1143.

Johnson, J. M., & Khoshgoftaar, T. M. (2021). Medical provider embeddings for healthcare fraud detection. *SN Computer Science*, *2*(4), 1–15.

Johnson, M. E., & Nagarur, N. (2016). Multi-stage methodology to detect health insurance claim fraud. *Health Care Management Science*, *19*(3), 249–260.

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., et al. (2015). Using data mining to detect health care fraud and abuse: a review of literature. *Global Journal of Health Science*, *7*(1), 194.

King, K. M. (2014). *Medicare fraud, progress made, but more action needed to address medicare fraud, waste and abuse:.*. United States Government Accountability Office.

Kumar, M., Ghani, R., & Mei, Z.-S. (2010). Data mining to predict and prevent errors in health insurance claims processing. In *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 65–74).

Lucas, Y., Portier, P.-E., Laporte, L., He-Guelton, L., Caelen, O., Granitzer, M., et al. (2020). Towards automated feature engineering for credit card fraud detection using multi-perspective HMMs. *Future Generation Computer Systems*, *102*, 393–402.

Luo, W., & Gallagher, M. (2010). Unsupervised DRG upcoding detection in healthcare databases. In *2010 IEEE international conference on data mining workshops* (pp. 600–605). IEEE.

Matloob, I., Khan, S. A., & Rahman, H. U. (2020). Sequence mining and prediction-based healthcare fraud detection methodology. *IEEE Access*, *8*, 143256–143273.

McGhin, T., Choo, K.-K. R., Liu, C. Z., & He, D. (2019). Blockchain in healthcare applications: Research challenges and opportunities. *Journal of Network and Computer Applications*, *135*, 62–75.

Moshagen, M. (2010). Multitree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, *42*(1), 42–54.

Mubarakali, A., Bose, S. C., Srinivasan, K., Elsir, A., & Elsier, O. (2019). Design a secure and efficient health record transaction utilizing block chain (SEHRTB) algorithm for health record transaction in block chain. *Journal of Ambient Intelligence and Humanized Computing*, 1–9.

Naidoo, K., & Marivate, V. (2020). Unsupervised anomaly detection of healthcare providers using generative adversarial networks. *Responsible Design, Implementation and Use of Information and Communication Technology*, *12066*, 419.

Ormerod, T., Morley, N., Ball, L., Langley, C., & Spenser, C. (2003). Using ethnography to design a mass detection tool (MDT) for the early discovery of insurance fraud. In *CHI'03 extended abstracts on human factors in computing systems* (pp. 650–651).

Ortega, P. A., Figueroa, C. J., & Ruz, G. A. (2006). A medical claim fraud/abuse detection system based on data mining: A case study in Chile. *DMIN*, *6*, 26–29.

Ozbayoglu, A. M., Gudelek, M. U., & Sezer, O. B. (2020). Deep learning for financial applications: A survey. *Applied Soft Computing*, *93*, Article 106384.

Pflaum, B. B., & Rivers, J. S. (1991). Employer strategies to combat health care plan fraud. *Benefits Quarterly*, *7*(1), 6.

Priya, S., & Uthra, R. A. (2021). Comprehensive analysis for class imbalance data with concept drift using ensemble based classification. *Journal of Ambient Intelligence and Humanized Computing*, *12*(5), 4943–4956.

Sahmoud, S., & Topcuoglu, H. R. (2020). A general framework based on dynamic multi-objective evolutionary algorithms for handling feature drifts on data streams. *Future Generation Computer Systems*, *102*, 42–52.

Sander, J., Ester, M., Kriegel, H.-P., & Xu, X. (1998). Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data Mining and Knowledge Discovery*, *2*(2), 169–194.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems*, *42*(3), 1–21.

Shanmugapriya, E., & Kavitha, R. (2019). Medical big data analysis: preserving security and privacy with hybrid cloud technology. *Soft Computing*, *23*(8), 2585–2596.

Simborg, D. W. (2008). Healthcare fraud: whose problem is it anyway? *Journal of the American Medical Informatics Association*, *15*(3), 278–280.

Singh, P., & Meshram, P. A. (2017). Survey of density based clustering algorithms and its variants. In *2017 international conference on inventive computing and informatics (ICICI)* (pp. 920–926). IEEE.

Sowah, R. A., Kuuboore, M., Ofoli, A., Kwofie, S., Asiedu, L., Koumadi, K. M., et al. (2019). Decision support system (dss) for fraud detection in health insurance claims using genetic support vector machines (gsvms). *Journal of Engineering*, *2019*.

Štefan, F., & Bajec, M. (2008). Holistic approach to fraud management in health insurance. *Journal of Information and Organizational Sciences*, *32*.

Thrun, M. C. (2021). Distance-based clustering challenges for unbiased benchmarking studies. *Scientific Reports*, *11*(1), 1–12.

van Capelleveen, G., Poel, M., Mueller, R. M., Thornton, D., & van Hillegersberg, J. (2016). Outlier detection in healthcare fraud: A case study in the medicaid dental domain. *International Journal of Accounting Information Systems*, *21*, 18–31.

Yamanishi, K., Takeuchi, J.-I., Williams, G., & Milne, P. (2004). On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining and Knowledge Discovery*, *8*(3), 275–300.

Yang, W.-S., & Hwang, S.-Y. (2006). A process-mining framework for the detection of healthcare fraud and abuse. *Expert Systems with Applications*, *31*(1), 56–68.

Zhang, Z., Chen, L., Liu, Q., & Wang, P. (2020). A fraud detection method for low-frequency transaction. *IEEE Access*, *8*, 25210–25220.

Zhou, S., He, J., Yang, H., Chen, D., & Zhang, R. (2020). Big data-driven abnormal behavior detection in healthcare based on association rules. *IEEE Access*, *8*, 129002–129011.