International Journal of Computers and Applications

ISSN: 1206-212X (Print) 1925-7074 (Online) Journal homepage: www.tandfonline.com/journals/tjca20

Enhancing healthcare data integrity: fraud detection using unsupervised learning techniques

Maithri Bairy, Balachandra Muniyal & Nisha P. Shetty

To cite this article: Maithri Bairy, Balachandra Muniyal & Nisha P. Shetty (2024) Enhancing healthcare data integrity: fraud detection using unsupervised learning techniques, International Journal of Computers and Applications, 46:11, 1006-1019, DOI: 10.1080/1206212X.2024.2408262

To link to this article: https://doi.org/10.1080/1206212X.2024.2408262

Published online: 09 Oct 2024.

INTERNATIONAL JOURNAL OF COMPUTERS AND APPLICATIONS
2024, VOL. 46, NO. 11, 1006–1019
https://doi.org/10.1080/1206212X.2024.2408262

Enhancing healthcare data integrity: fraud detection using unsupervised learning techniques

Maithri Bairy [a], Balachandra Muniyal [a,b] and Nisha P. Shetty[a]

[a]Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India; [b]MAHE-ISAC CoE for Cybersecurity, Manipal Academy of Higher Education, Manipal, India

ABSTRACT

Data in healthcare forms the backbone of any treatment and decision-making for patients. However, the data from a healthcare institution can sometimes be prone to abnormalities, thereby putting treatment and patient safety in jeopardy. This paper points to the dire need for reliable anomaly detection systems in the healthcare industry. It employs various unsupervised learning methods, including Isolation Forest, Local Outlier Factor (LOF), K-Nearest Neighbours (KNN), and autoencoder models for detecting abnormalities in healthcare data with better accuracy. Anomaly detection capabilities also allow health providers to reduce risks and provide some assurance of the integrity of the data, as these capabilities are more likely to indicate unusual patient profiles or incorrect test results. Isolation Forest, LOF, and KNN are the preliminary methods for performing anomaly detection in this work, with Isolation Forest yielding the best results. Then, autoencoder models that learn subtle variations and complex patterns in data are employed. This paper seeks to enhance anomaly detection in terms of accuracy and reliability to ensure better healthcare data quality and patient safety.

ARTICLE HISTORY
Received 14 June 2024
Accepted 13 September 2024

KEYWORDS
Anomaly detection; fraud detection; machine learning; deep learning; autoencoders; healthcare

1. Introduction

The healthcare industry's critical role as an application domain for statistical analysis cannot be overstated. However, despite its significance, a substantial portion of healthcare funding is lost annually to fraud, waste, and abuse (FWA), amounting to billions of dollars in programs like Medicaid and Medicare [1]. Healthcare data is highly dynamic and includes insurance claims, medical records, clinical data and provider data, all of which have posed a big challenge to traditional fraud detection algorithms. Besides, there has been an increase in the number and complexity of health data due to the rise in digital health systems and digitization of patient records.

For policymakers and healthcare providers, this abundance of data offers both opportunities and challenges. While digitization has undoubtedly enhanced patient care effectiveness and improved health outcomes, it has also created new avenues for fraudulent activities. Fraudsters take advantage of weaknesses in healthcare systems to run a range of scams, making fraudulent claims to tampering with patient data through identity theft, upcoding, or phantom billing. As healthcare fraud is pervasive and complex, it requires as much insight as possible into the motivations at play for all parties in this ecosystem of provision and payment [2].

Machine learning and deep learning algorithms have stepped in to represent an effective counterbalance against healthcare fraud, waste and abuse. As opposed to conventionally developed rule-based systems, machine learning algorithms grant adaptability and learning from extensive and complex datasets. The algorithms can reveal minute divergences from normal behavior that may represent fraud by making use of some of the most recent approaches including anomaly detection. Moreover, their versatility extends the ability to analyze in-depth the various sources of healthcare data like insurance claims, medical records and provider information.

Furthermore, machine learning algorithms possess the inherent capability to detect emerging fraud trends and adapt to evolving strategies employed by fraudsters. This adaptability is crucial in the constantly evolving landscape of healthcare fraud, where new tactics and schemes continually emerge [3]. Additionally, healthcare organizations can embed various machine learning algorithms into the fraud detection systems to automate the detection processes without having high dependence on subject matter experts and manual intervention. This will further enable them to allocate more resources effectively to the fraud prevention and mitigation techniques along with enhancing the fraud detection efforts to become more effective and scalable. By applying the power of machine learning algorithms in anomaly detection, this research seeks proactive detection and prevention of fraudulent activities in healthcare data with a view to protecting integrity in health programs and averting or, at best, reducing economic losses.

The efficient application of machine learning algorithms to anomaly detection can be considered a milestone in fraud detection in healthcare. This paper seeks to utilize the potential of machine learning algorithms in analyzing complex healthcare datasets, identifying fraudulent patterns from them and adapting to emerging threats as part of continuous contributions to the fight against fraud, waste and abuse in healthcare systems.

2. Problem definition

The various complications inherent in patient-oriented healthcare data, such as insurance claims, medical records, clinical data, and provider information, act as a barrier in the way of effective detection of healthcare fraud. The nature of healthcare data presents a plethora of challenges in fraud detection. First, because there are so

CONTACT Balachandra Muniyal bala.chandra@manipal.edu Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India; MAHE-ISAC CoE for Cybersecurity, Manipal Academy of Higher Education, Manipal, India

many sources of data coming from different corners, inconsistencies and anomalies indicative of fraud may become hard to identify. This leads to the problem of fragmented and scattered healthcare data across systems and platforms, for example, its detailed analytics and integration. Fraud schemes in healthcare are getting larger and more agile, taking advantage of the loopholes in today's detection infrastructure. Since fraudsters use many various kinds of strategies, including identity theft, upcoding, and submitting invoices for services never provided, it is hard to recognize real fraud by using traditional rule-based procedures.

Moreover, fraud in healthcare usually involves a conspiracy on the part of several parties, including patients, healthcare providers, and third-party organizations. Traditional fraud detection techniques can't handle the complexity involved in the analysis of key relationships and patterns for collusion-based fraud detection amongst various sources. Also, the nature of healthcare data being dynamic contributes to more challenges for real-time fraud detection. Traditional methods for detecting fraud may not find newly emerging fraud trends and patterns, as their logic is based on predefined rules and history. Another challenge in the area of fraud detection in healthcare is the presence of so-called false positives where real transactions are flagged as fraud. Such false positives take more time by initiating groundless investigations and consumption of resources, consequently delaying processes of fraud detection.

To make matters worse, the regulatory landscape that surrounds the detection of healthcare fraud is complex and constantly changing. The many regulations and compliance mandates which also apply to the operations of fraud detection in healthcare organizations make the operation of fraud detection even more difficult. In essence, healthcare fraud detection is a rather knotty issue. Innovative solutions are needed as a way to overcome obstacles caused by the dynamic and complex nature of the data present.

3. Related works

The study by Branting et al. [1] estimated the probability of healthcare fraud (HCF) using novel network algorithms and graph analytics. The research focused on behavioral similarity calculations using healthcare activities and predicted risk transmission through geospatial collocation, by using free source datasets. The algorithms demonstrated a high degree of accuracy when evaluated on exclusion prediction, highlighting the importance of characteristics associated with the spread of risk. In order to further improve predictive accuracy, the article suggests a richer graph representation and a wider range of targets. This suggests that diverse behavioral similarity measures could lead to advancements in the assessment of HCF risk.

To uncover strategies for identifying healthcare fraud, Thaifur et al. [2] carried out a systematic review. They searched databases such as PubMed, Wiley, ScienceDirect, and Google Scholar for papers using keywords. According to nine publications that were analyzed, medical professionals, particularly physicians, predominate the fraud cycle when it comes to filing false claims and multiple insurance claims. Information technology specialists play a crucial part in fraud detection as secondary data tracking emerged as the most generally used strategy. The article recommends more research to find flaws in fraud detection techniques used worldwide, which would help develop national and professional legal penalties for fraud.

Liu et al. [3] applied graph-analysis approaches to handle the difficult problem of identifying fraud, waste, and abuse (FWA) in health-care data. They saw every health-care data collection as a heterogeneous network and used graph analysis to find linkages, persons, changes over time, anomalies in geography, and network structures that could be of concern. The network explorer visualization interface allowed users to filter and zoom into data details. The technique had been used on multiple government and commercial websites, effectively detecting overpayments amounting to millions of dollars each month. Upcoming initiatives include integrating user feedback, streamlining algorithms for real-time claim stream analysis, and making configuration simple for a variety of data types.

Agrawal et al. [4] addressed the escalating issue of fraud in the healthcare sector, causing a surge in healthcare expenditure. Using the Class Weighing Scheme (CWS) and Adaptive Synthetic Oversampling (ADASYN) data balancing approaches, their study focused on a comparison of several machine learning models. The study emphasized the need to combat fraudulent situations by incorporating data preprocessing and exploratory data analysis (EDA). In addition to offering insights into fraud detection in healthcare insurance systems and making suggestions for prospective advancements utilizing deep learning models for future study, the proposed technique attempts to validate the effectiveness of various machine learning models.

Dornadula et al. [5] contributed to the ongoing discourse on credit card fraud detection. The research suggested a novel technique involving behavioral pattern extraction by clustering based on transaction amounts in response to the growing hazards associated with online transactions. The study applied different classifiers to different client groups using a sliding window technique and used a feedback mechanism to address idea drift. Their analysis of the European credit card fraud dataset illustrated the usefulness of Matthews Correlation Coefficient and tackled dataset imbalance through approaches like SMOTE. Robust algorithms for fraud prediction include random forest, decision tree, and logistic regression.

The vulnerability of the healthcare industry to financial fraud, specifically credit card theft, is discussed in the paper by Mehbodniya et al [6]. The authors investigated machine learning and deep learning techniques as a result of the difficulties in detecting fraud posed by the ongoing expansion of electronic payments. The accuracies of the Naive Bayes, Logistic Regression, KNN, Random Forest, and Sequential Convolutional Neural Network methods were compared. KNN performed surprisingly better than other approaches, suggesting that more research is necessary. The report suggested enhancements for improved credit card transaction fraud detection in the healthcare industry, including transfer learning and hyperparameter tuning.

The effectiveness of machine learning, especially when applying the Logistic Regression and Support Vector Machine algorithms, in detecting payment fraud has been demonstrated by recent studies. Oza's study [7], which focused on the Paysim dataset, showed good accuracy and few false positives, particularly in TRANSFER transactions. The suggested class weight-based method worked well for addressing the problem of unbalanced datasets. Future possibilities for improving approaches include investigating Paysim as a time series, creating user-specific models based on transactional history and integrating decision trees for categorical data. The goal of these developments is to improve fraud detection, which is important for making business decisions in the world of digital payments.

Raghavan and Gayar [8] assessed machine learning and deep learning techniques, such as CNN, RBM, DBN, KNN, Random Forest, SVM, and autoencoders, in their fraud detection study. The metrics AUC, MCC, and Cost of Failure are evaluated using datasets from Germany, Australia, and Europe. They emphasized the recent deployment of deep learning models in their empirical analysis, which span more than 20 years of fraud detection research. For bigger datasets, the study suggested using SVMs in conjunction with CNNs; for smaller datasets, ensemble techniques (SVM, Random Forest,

KNN) are advised. While acknowledging its limitations, the study highlighted how adaptable Autoencoders are in dynamic situations, offering useful information for businesses and practitioners.

Raman et al. [9] investigated cyber security fraud detection using machine learning in their work. Using data from the EU, Canada, and the Netherlands, they assessed well-known machine learning techniques such as KNN, RF, SVM, and deep neural networks (DBN), which include autoencoders, classifiers, multiple solutions, and DBN. The cost of failure, AUC, and MCC are examples of evaluation metrics. The paper highlighted the effectiveness of machine learning algorithms and the current adaptability of deep learning models, drawing on more than 20 years of research on fraud detection. It recommended CNNs for better outcomes, SVMs for bigger datasets, and iterative techniques for dynamic environments.

The study by Yoo et al. [10] addressed the critical issue of Medicare fraud detection through a pioneering comparative analysis of machine learning and Graph Neural Networks (GNN). With an emphasis on the connections between medical providers, beneficiaries, and physicians, the study applied graph analysis to tabular information to improve the accuracy of fraud detection. In this research, graph centrality measurements are used to introduce GNN models and regular ML models. The latter model is shown to have better recall and F1-score. The novel method stressed on how important graph centrality is for capturing intricate fraud links. The paper made a significant contribution by providing effective fraud detection techniques, which are essential for healthcare insurance operations, highlighting the superiority of graph-based ML over GNNs, and providing insightful information for further investigation into fraud detection models.

A framework for detecting anomalies in Medicare claims was developed by Kemp et al. [11]. Their study tackled the requirement for higher rates of non-compliant activity detection in medical claims from Australian providers. The investigation of novel approaches is prompted by the fact that current procedures lag behind international benchmarks. Their strategy comprised putting into practice three workable techniques intended to improve interpretability, context recognition, and cost recovery. The study intended to improve detection rates and reduce fraudulent and wasted claims by using novel anomaly detection approaches, hence supporting the sustainability and equity of healthcare initiatives.

A thorough investigation on machine learning-based methods for healthcare fraud detection was carried out by Waghade and Karandikar [12]. Their study highlighted how fraud in the healthcare sector is becoming a bigger issue, especially when it comes to the abuse of medical insurance programs. They emphasized the significance of sophisticated machine learning techniques, such as supervised, unsupervised, and semi-supervised learning approaches, for boosting fraud detection and raising the efficacy and affordability of healthcare systems by analyzing a variety of research in the literature.

Medicare fraud detection using random forest with class imbalanced large data was studied by Bauder and Khoshgoftaar [13]. Their research emphasized how crucial it is to use big data–such as provider payments and patient records to effectively detect fraud in the healthcare sector. They showed that the widely used 50:50 balanced ratio does not always produce the best fraud detection results by evaluating various class distributions and using random undersampling approaches. Their findings suggested that the 90:10 class distribution is the best option for fraud detection performance and highlighted the necessity of using suitable sampling approaches to overcome class imbalance.

A novel graph network technique for modeling and analyzing patient flow in hospital emergency departments (EDs) was presented by Reychav et al. [14]. To gain insights from the system, they computed metrics like degree centrality and shortest pathways using a time-varying graph (TVG) technique kept in a Neo4j database. Their research demonstrated the usefulness of TVG modeling in illuminating the dynamic relationships between hospital departments and consultants, giving administrators insightful information for allocating resources and making decisions. To determine total patient satisfaction, they also examined customer ratings and reviews. This illustrated how knowledge graphs and sentiment analysis may be used to improve ED procedures and patient care results.

The study by Pourhabibi et al. [15] developed a framework to synthesize the literature on the application of graph-based anomaly detection (GBAD) methods in fraud detection published between 2007 and 2018, taking into account the many GBAD approaches presented for fraud detection. GBAD is a popular method for examining connectivity patterns in communication networks and spotting suspicious activity. In order to strengthen the technique's credibility, this study looked at current trends and pinpointed the major issues that need further investigation. The shortcomings that have been uncovered encourage data scientists to conduct additional empirical study in this area. This work also provided practitioners with a road map to understand how their network's characteristics, various anomalies, and suitable graph-based techniques align with their requirements and use cases.

Together with data ingestion techniques, the work by Ali and Praveen [16] aimed to synthesize various strategies for outlier identification into a systematic and generalized description for a given set of data. This was accomplished by using machine learning algorithms, such as autoencoders, KNN, MCD, etc., with appropriate parameter setting, to transform the data. When utilizing various algorithms for anomaly/outlier detection, the parameters were fine-tuned based on the provided data set and were even scored in relation to performance comparison. Additionally, two significant algorithms- Isolation Forest and Auto encoders were compared, and conclusions were formed. The prior preparation of the data was also carried out.

In the work by Munir et al. [17], 13 different anomaly detection techniques were analyzed using two popular streaming data sets. Four distinct evaluation measures were employed to assess the strategies from various angles. The deep learning-based anomaly detection techniques outperformed the conventional anomaly detection techniques, according to these findings. The results demonstrated that, using the Yahoo Webscope data set, deep learning-based anomaly detection techniques outperformed other techniques in the majority of evaluation criteria. When it came to temporal complexity, PCA outperformed deep learning-based techniques.

In the study by Bowie et al. [18], ongoing research on the application of cutting-edge anomaly detection methods based on Long-Short-Term-Memory (LSTM) neural networks to medical data was conducted. Given that LSTMs can preserve information regarding both immediate and long-term dependencies on the model outcome, it was postulated that they offer a flexible and robust method to temporal anomaly detection in medical data sets. LSTM-based methods were said to perform better than popular statistical techniques like control charts and regression.

The research by Reddy et al. [19] suggested a new approach for identifying outliers from various medical datasets. Taking into account that medical data analyze health issues, the suggested method operated on the foundation of both supervised and unsupervised learning. The outliers in medical data were detected by this algorithm. The efficiency of both local and global data factors for real-time outlier detection in medical data was estimated. Whichever model was employed in this instance, it was trained and

tested using medical data. Research is carried out using integrated medical datasets. The statistical findings demonstrated that the outlier recognition technique based on machine learning gave the best accuracy.

An effective autoencoder-based model for anomaly detection in cloud computing networks was presented in the paper by Torabi et al. [20]. Reconstruction error served as an anomaly or classification metric. Apart from distinguishing anomalous data from typical data, research was done on grouping different kinds of anomalies. It is believed that the reconstruction error is a vector. This allowed their model to use each input feature's reconstruction error as a classification measure. The analysis demonstrated that the accuracy, recall, false-positive rate, and F1-score metrics-wise performance of the suggested strategy had significantly improved over the previous ones.

In the paper by Sattarov et al. [21], a denoising autoencoder framework for mixed type tabular data was provided to explain anomalies. The method was especially targeted at anomalies that are wrong observations. To do this, each sample column (cell) containing possible errors was localized, and associated confidence scores were assigned. To correct the mistakes, the model also offered predicted cell value estimations. When used for this purpose, denoising autoencoders already performed better than alternative methods in terms of expected value rates and cell error detection rates. For each cell entry, the framework generated confidence scores of probable inaccuracies and suggested corresponding estimated values to correct the faults. Furthermore, an improved extension that makes use of an extended loss that has been tailored for cell error detection is suggested.

A thorough analysis of autoencoders was presented by Berahmand et al. [22], commencing with an explanation of the fundamental ideas behind traditional autoencoders and how they were developed. Next, a taxonomy of autoencoders according to their architectures and guiding principles was taken, examined and the relevant models were elaborated upon in detail. Furthermore, how autoencoders were used in a variety of domains, including speech recognition, machine learning, natural language processing, complex networks, recommender systems, and anomaly detection was looked into.

Autoencoder and DCGAN are the two types of anomaly detectors that were trained in the study by Kopčan et al. [23]. Three different dataset types were used to train them: MNIST, Fashion-MNIST, and CIFAR-10. Given the frequency of anomaly occurrence for DCGAN networks and autoencoders, the shared optimal number of latent variables and the optimal decision threshold for autoencoders was found. Under an anomaly expectation probability of 0.5, the anomaly detection errors for the two detectors trained on the most basic database, MNIST, were 0.08% (AAE) and 1.89% (DCGAN).

The paper by Gonzalez et al. [24] offered a possible strategy for tracking people's daily activities in the least invasive way possible. In this instance, everyday activity patterns that exhibit anomalies were identified through analysis of domestic appliance utilization. Based on an autoencoder architecture, a neuronal model for the identification of aberrant behavior was put forth. The potential gains were examined by contrasting this approach with a variational autoencoder. The idea was validated by using the widely recognized dataset known as UK-DALE. This approach was a non-intrusive means of providing care for individuals with dementia or Alzheimer's disease since it could accurately identify circumstances in which the user needs assistance and helped steer clear of potentially hazardous scenarios brought on by these conditions.

In the study by Alhassan et al. [25], a predictive Deep Learning model was developed to assess in-hospital patient mortality risk. The King Abdullah International Research Centre (KAIMRC) is a unique time-stamped dataset that is naturally unbalanced and was used to train the stacked denoising autoencoder (SDA). The work was contrasted with popular deep learning techniques that use various data balance strategies. With an accuracy of 77.13% for the Recall macro, the suggested model was shown to beat popular deep learning techniques and solved the issue of imbalanced data.

The study by Sakurada and Yairi [26] showed that autoencoders can identify minor anomalies that linear PCA was unable to pick up on. Moreover, autoencoders could be extended to denoising autoencoders, which would improve their accuracy. Additionally, autoencoders could be helpful as nonlinear methods without requiring the kind of intricate calculation that kernel PCA did. Upon analyzing the learnt features in the autoencoders' hidden layer, the authors inferred that autoencoders effectively learn the normal state and react differently to anomalous input.

An unsupervised anomaly detection method called RULEAD was introduced in the study by Senaratne et al. [27], which found anomalous records in datasets. Several features were first extracted from attributes, then these learn normality to derive normal and abnormal feature vectors, rank and select the most distinctive features, cluster using the selected features, and finally provide a symbolic depiction of the decision rules underlying the anomalies noticed. Domain experts can use these rules to learn more about the records in a graph. RULEAD was capable of comparing and connecting several properties by taking them into account simultaneously.

Biomedical data are frequently gathered as time series, pictures and electronic health records. When compared to a single encoder, an ensemble of autoencoders (EoAE) can offer better detection; nevertheless, the effectiveness of this approach can vary depending on a number of parameters, such as the variety of the generated data, the accuracy of the individual AEs, and the combining of their results. The paper by Nawaz et al. [28] did a thorough narrative literature analysis on the application of EoAEs to various biomedical data types. By utilizing the advantages of several AEs, such an ensemble it offered a viable method for anomaly detection in biomedical data and presented the possibility of performance enhancement.

The study by Mues et al. [29]can be utlized by researchers to get an overview of Medicare data, including what kinds of data are collected and how they might be used for epidemiologic and health outcomes research. The design of a study using Medicare data highlighted the advantages, drawbacks, and important factors. Furthermore, the effects that policy changes from the Centres for Medicare and Medicaid Services (CMS) might have on data collecting, coding, and eventually data-derived findings were examined.

By identifying outlying features, the work by Samariya et al. [30] seeked to both swiftly and effectively identify anomalies and provide an explanation for their classification as anomalies. Initially, four methods for detecting anomalies and algorithms for mining outlying aspects were examined. Next, the most effective anomaly detection algorithm was selected after an analysis of the performance of several anomaly detection strategies. Following that, the outlying features of the top k anomalies were found after they qualify as a query. Performance was compared on sixteen real-world healthcare datasets. The experimental findings demonstrated that SiNNE, the most recent isolation-based outlying aspect mining measure, performed exceptionally well and showed promise in this task.

Healthcare fraud is a significant financial burden in the United States, impacting insurance costs and patient well-being. The paper by Kumaraswamy et al. [31] explored current literature on healthcare fraud detection, outlined methods and discussed the complexities of integrating data and adapting models. It underscored the importance of ongoing research to bridge gaps in applying these systems to practical healthcare environments, suggesting ways to improve detection reliability and effectiveness.

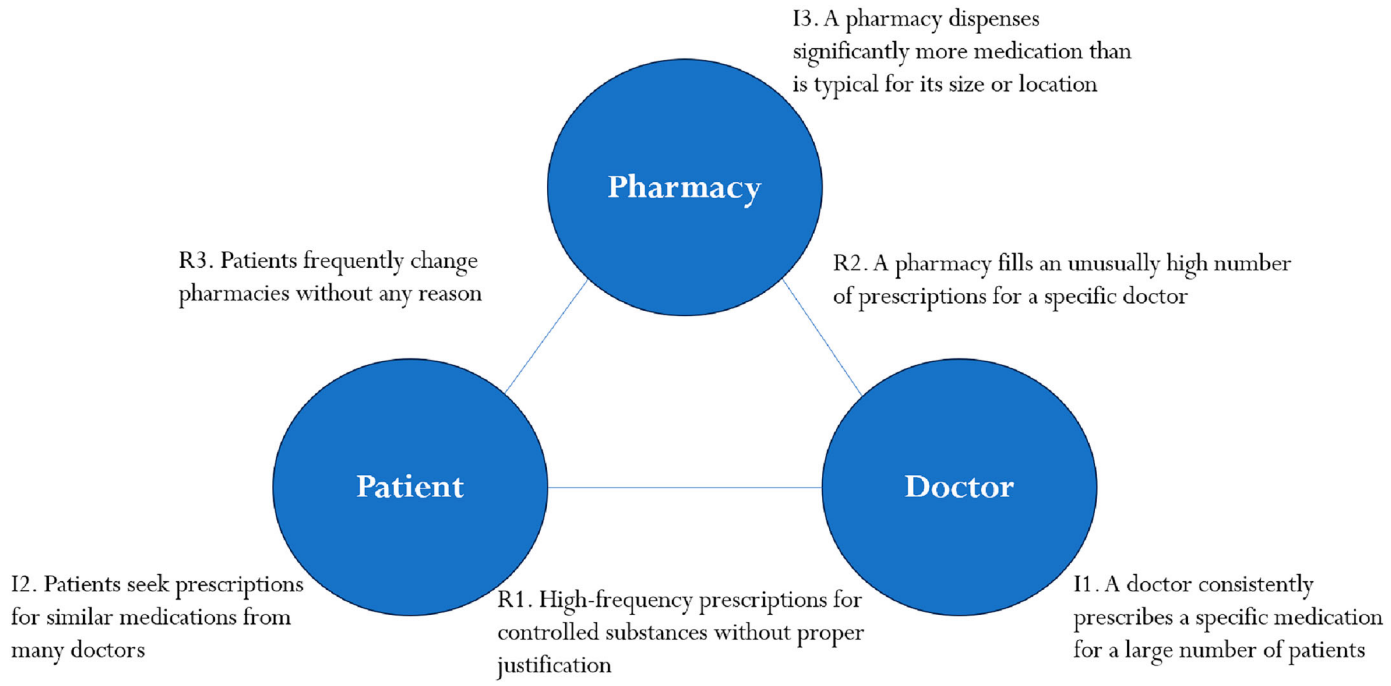


Figure 1. Anomalies in patient data graph.

Yu et al. [32] suggested the DWAD-LDVP method in their article, which integrated a verification model with an adaptive dynamic sliding window mechanism. The technique improved anomaly identification accuracy over state-of-the-art methods by converting local density values, as determined by the vector dot product, into outlier scores. In addition, it minimized needless computational waste by processing data streams effectively using an incremental computing module. The superior performance of DWAD-LDVP are proven as it exhibited competitive results in terms of accuracy, precision, recall, F1 score, ROC, and AUC.

4. Methodology

Healthcare data fraud can occur in many different ways. Some common fraudulent activities seen in most data records are depicted in Figure 1.

4.1. Data preprocessing and feature engineering

An initial exploration of the dataset to have a deeper look into the available features and check their relevance in fraud detection is carried out. The dataset has 27 relevant features, out of which 10 are numerical and others are categorical. Then, each feature is filtered for missing values and duplicate rows during data cleaning. The datatypes of the features are identified. The handling of missing data included imputation using mode values for features of a categorical nature and mean values for features of a numerical nature. In order to perform the correlation analysis, all features needed label encoding. Next, feature engineering is performed. Figure 2 shows the correlation heatmap for all 27 features. Attributes with correlation of more than 0.9 were removed in order to avoid multicollinearity. This resulted in a total of 24 features to be used with the training models. While mode and mean imputation methods were employed in this research, it has to be taken into consideration that other imputation strategies could yield different results. For example, median imputation or model-based imputation methods can provide other insights and may affect the robustness depending on the nature of the missing data.

4.2. Training anomaly detection machine learning models

For training, we opt for three distinct algorithms: Isolation Forest, Local Outlier Factor (LOF), and K-Nearest Neighbors (KNN), each offering unique advantages in detecting anomalies.

One famous anomaly detection method that is well-known for its efficacy and efficiency in locating outliers in datasets is the Isolation Forest algorithm. It functions according to the idea that anomalies are areas of data that are separated from the bulk of the data. This approach works very well with high dimensionality datasets that have intricate feature interactions. Fundamentally, the Isolation Forest algorithm builds an isolation tree random forest. To construct a tree, a collection of features is chosen at random, and the data is divided along these features until isolated instances are found. The main idea behind this method is that since anomalies are naturally distinct from regular occurrences and tend to be found in sparser areas of the feature space, they should be easier to isolate with fewer splits.

An effective method for locating outliers in datasets is the Local Outlier Factor (LOF) algorithm, which takes into account the density of data points in the outlier's immediate vicinity. As LOF concentrates on the local structure of the data rather than the global structure, it is especially useful for datasets with different densities or clusters [32]. Fundamentally, each data point's anomaly score is calculated by the LOF algorithm using its local density in relation to that of its neighbors. Higher anomaly scores are given to instances that have a much lower density than those of their neighbors, which makes them stand out. This method enables LOF to detect abnormalities even in areas with low overall density by adapting to the local structure of the data.

A popular method for detecting anomalies is the K-Nearest Neighbours (KNN) algorithm, which uses the closeness of data points in the feature space as its basis [8]. By calculating the separation between each data point and its k-nearest neighbors and

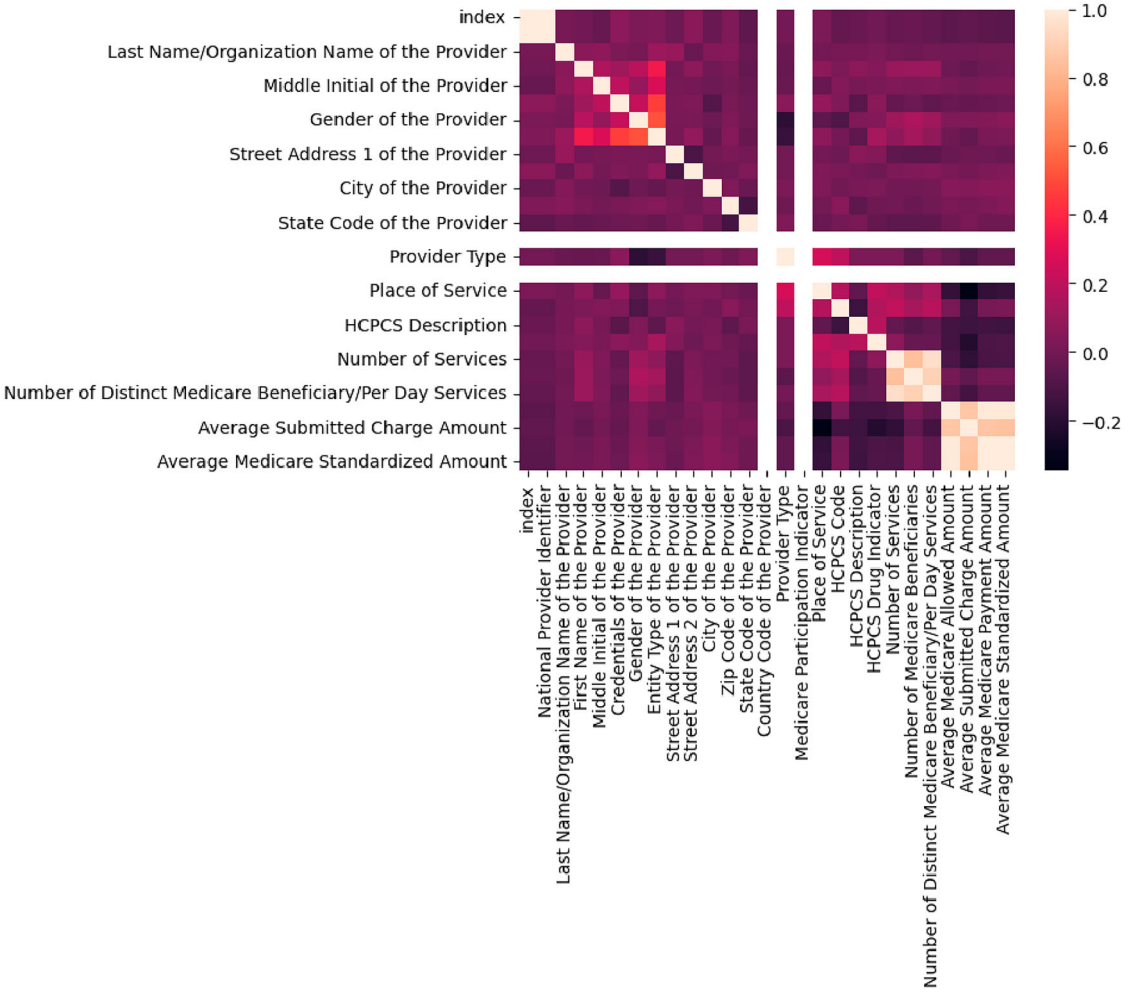


Figure 2. Correlation heatmap of 27 features.

allocating anomaly scores based on these distances, it finds outliers. Fundamentally, the KNN algorithm compares data points according to their feature values to determine how similar they are [6]. Higher anomaly scores are given to instances that differ greatly from their neighbors in terms of feature values, which classify them as outliers. With this method, KNN can identify abnormalities in datasets with intricate feature interactions and a range of densities [9].

The major advantage of the Isolation Forest algorithm is its effective detection of anomalies without relying on density estimation or distance measures [16]. Unlike other methods, it utilizes the natural structure in data so that anomalies can be found in fewer stages. Due to this, Isolation Forest is a good method for finding outliers in big datasets with diverse densities and complicated distributions. Besides, Isolation Forest can handle datasets with diverse types of attributes – numerical and categorical variables and is resistant to noise introduced by redundant features. Further, on account of its versatility, it could be applied for several healthcare datasets because many of them consist of heterogeneous information, such as patient demographics, clinical records and billing codes. A key advantage of the LOF algorithm is that it captures complex, even nonlinear relationships between features in the data. LOF can be used as a method for anomaly detection in healthcare data, which very often exhibits complex patterns and correlations [30]. The KNN algorithm is beneficial due to its simplicity and ease of use. The algorithm can be applied to datasets that involve a mix of features and those that require no assumptions about the underlying distribution of data. Since healthcare data often combines several types of information, KNN provides a flexible technique for detecting anomalies.

4.3. Assessing machine learning models' performance

After training, the models are used to predict the Anomaly labels on the healthcare dataset. These labels are also assigned back to the original dataset for clarity. The models' performance is assessed based on the number of records that have been flagged as fraudulent after they have been tested on healthcare data. The contamination in each model is fixed at 0.05. Furthermore, to see the distribution of data points in a lower-dimensional space, TSNE (T-distributed Stochastic Neighbour Embedding) plots are produced. These charts help evaluate how well the models recognize outliers and shed light on how normal and anomalous occurrences cluster. The t-SNE plots and fraud detection capabilities of each model are used to compare how well the three models perform. The model chosen as the recommended anomaly detection model for healthcare data is the one with the highest percentage of total anomalies found.

4.4. Classification on the labeled healthcare data

After implementing and evaluating various anomaly detection models on the healthcare dataset, further assessment consists of assessing the best model's efficiency by using a decision tree classifier. This decision tree classifier is trained on labeled data

obtained from results of the best-performing anomaly detection model.

Before the training of the decision tree classifier, we split the labeled dataset obtained from the results in the best-performing model in the prior stage into a training set and a test set in a 7:3 split. This will ensure that both normal and anomalous instances are similarly distributed in the training and test sets and hence give an unbiased evaluation of the classifier's performance. Stratified 10-fold cross-validation is used at the training phase to avoid overfitting the model and provide a more realistic performance assessment of the decision tree classifier. The stratification splits all the training data into 10 equal-size folds while maintaining the proportion between normal and anomalous instances in each fold. The classifier is trained 10 times. The decision tree classifier is trained on the training data after the data split and cross-validation approach is implemented.

Decision trees are popularly used for classification problems because of their interpretability and their ability to capture complex relationships between features. The algorithm for training a decision tree recursively divides the feature space into partitions or regions based on the values of the input features in a manner that maximizes the purity in the resulting subsets. It is performed until some stopping criterion, such as a maximum depth or minimum number of samples in a leaf node, is reached. Based on the labeled data, the decision tree learns to divide the feature space into regions that effectively distinguish normal cases from anomalies during training.

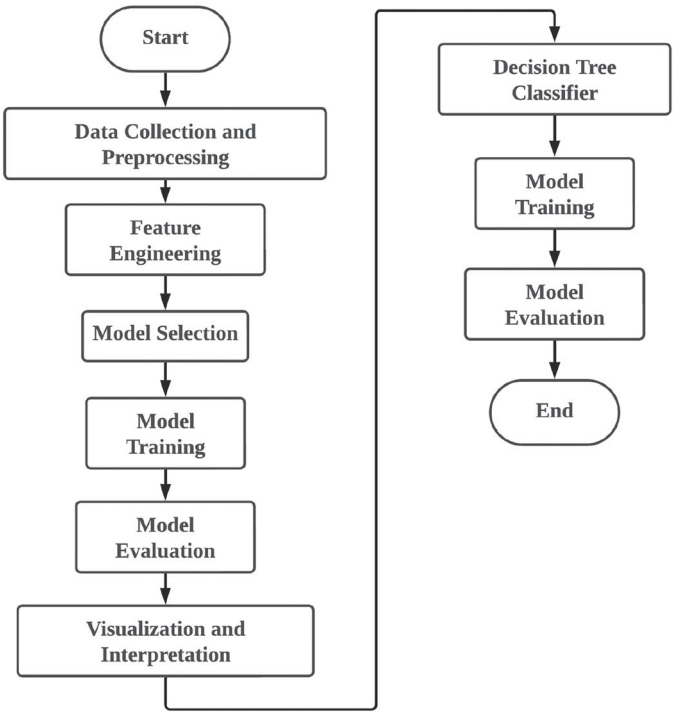


Figure 3. Methodology flowchart.

4.5. Assessing classifier's performance

After training the classifier on the test data, its performance is measured in terms of accuracy, precision, recall, and F1 score. This gives insight into how well the classifier can classify instances as normal or anomalous. Accuracy gives the overall percentage of correctly classified instances out of the total number of instances in the dataset. Precision is defined as the ratio of true positive predictions (correctly classified anomalies) out of all the instances that were predicted as anomaly. Recall is the ratio of true positive predictions out of all actual anomalies in the dataset. The F1 score is the harmonic average of precision and recall [15]. The confusion matrix shows the percentage of true positive, true negative, false positive, and false negative predictions, which are useful in knowing classifier efficiency.

Overall flow diagram till this point is depicted in Figure 3.

4.6. Using autoencoders

Autoencoders represent a neural network architecture useful in data compression, dimensionality reduction and unsupervised learning. An autoencoder learns a compressed form of input data by encoding input to its lower-dimensional representation after which the representation is decoded to the original input.

An encoder, a hidden layer and a decoder are the three primary parts of an autoencoder [18]. The encoder maps the input data to a lower-dimensional representation while the decoder maps a lower-dimensional representation back onto the original input data. The hidden layer or code is the lower dimension representation of the input data. By learning to minimize the difference between the input and the reconstructed output during training, the network is encouraged to learn a compressed representation of the input. Perhaps one of the biggest advantages with autoencoders is that they can uncover meaningful patterns in the data, including non-linear relationships among features. They can be used in many different domains like generating new data, anomaly detection and image compression [22].

In the case of anomaly detection, autoencoder training is done to learn a compressed representation of normal data. It should learn the normal patterns and relationships in a low-dimensional representation so that anomalies can be identified as data points that do not fit the pattern. Once the autoencoders are trained on normal data, they would be used to encode fresh data points where their reconstruction error can be compared with threshold value and in case reconstruction error exceeds the predetermined level, the data point is declared abnormal. Thus, this approach of using autoencoder enables the detection of abnormality in the new data in absence of explicit labeling of anomalies in training data [23].

One or more fully connected layers make up the encoder, which transforms the input data into a lower-dimensional representation. As one might expect fewer nodes in the hidden layer than either the input or the output, the network needs to learn how to represent the input in a compressed form. The network can thus learn non-linear relationships in the input data. In general, any nonlinear activation function can be employed here, like a rectified linear unit or sigmoid [24].

Usually a mirror copy of the encoder, the decoder consists of one or more connected layers that restore the original input space from the compressed representation. To create a reconstruction of the input data, the decoder's output layer needs to have an equal number of nodes as its input layer. The activation function utilized in the encoder and decoder is identical [25, 26].

These algorithms generate anomaly scores and an anomaly flag as an output and a general note is that the higher the anomaly score, the more chances of the data point to be anomalous [20]. Outliers are identified by large differences between input and reconstructed data.

In trial 2, the hidden neurons were one of the factors which were supposed to be changed along with the number of hidden layers when in search of the correct quantity of anomalies which is compatible with the quantity of data too.

Figures 4 and 5 depict the loss values vs number of epochs graphs for trial 1 and trial 2 respectively. Training loss curve is found to have a greater slope initially than the validation loss curve in loss values vs number of epochs graphs. The gap between the two curves in the

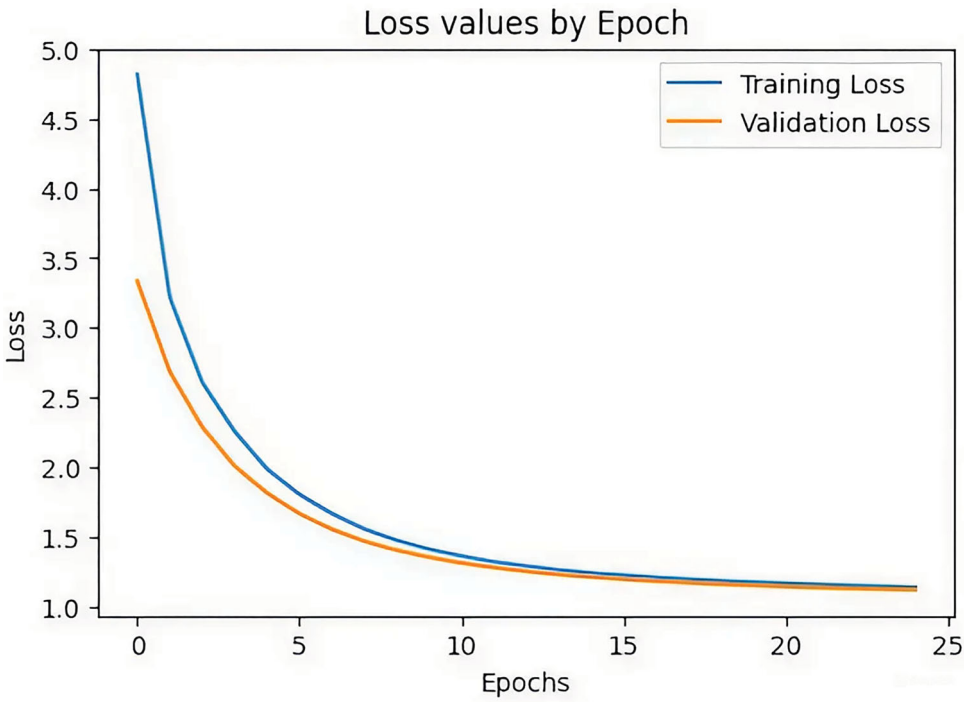


Figure 4. Loss values vs number of epochs for trial 1.

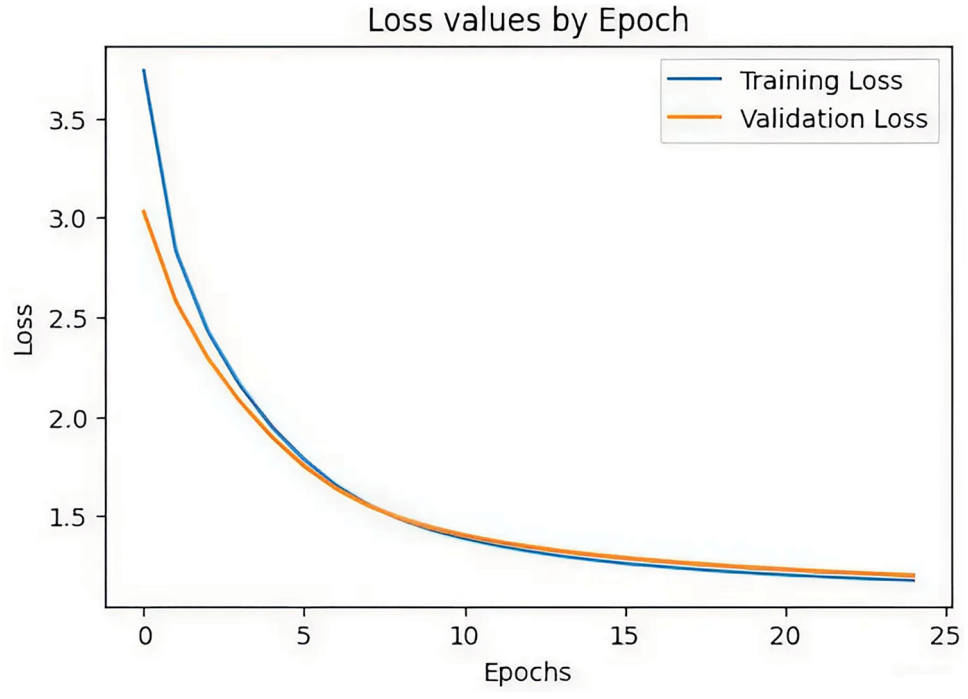


Figure 5. Loss values vs number of epochs for trial 2.

loss values vs number of epochs graphs starts to gradually decrease with the number of epochs increasing. This shows that the model is a good fit for the task of anomaly detection.

The Pyod autoencoder is trained on the CMS dataset. The selected autoencoder is sequential. During training, the autoencoder learns to reconstruct the input data while minimizing the reconstruction error. The architecture depicted by hidden neurons specifies how the input data is compressed into a lower-dimensional latent space and then reconstructed back to its original dimensionality.

During testing, anomalies are identified based on the reconstruction error. Data points that are poorly reconstructed (i.e. have high reconstruction error) are considered anomalies. The calculated values are mean_score, std_score, z-score, margin of error and threshold.

The z-score standardizes the data and determines just how unusual or typical a given data is within a data distribution. A z-score of 0 means that the value of the data point is exactly equal to the mean, while positive and negative z-scores determine values above

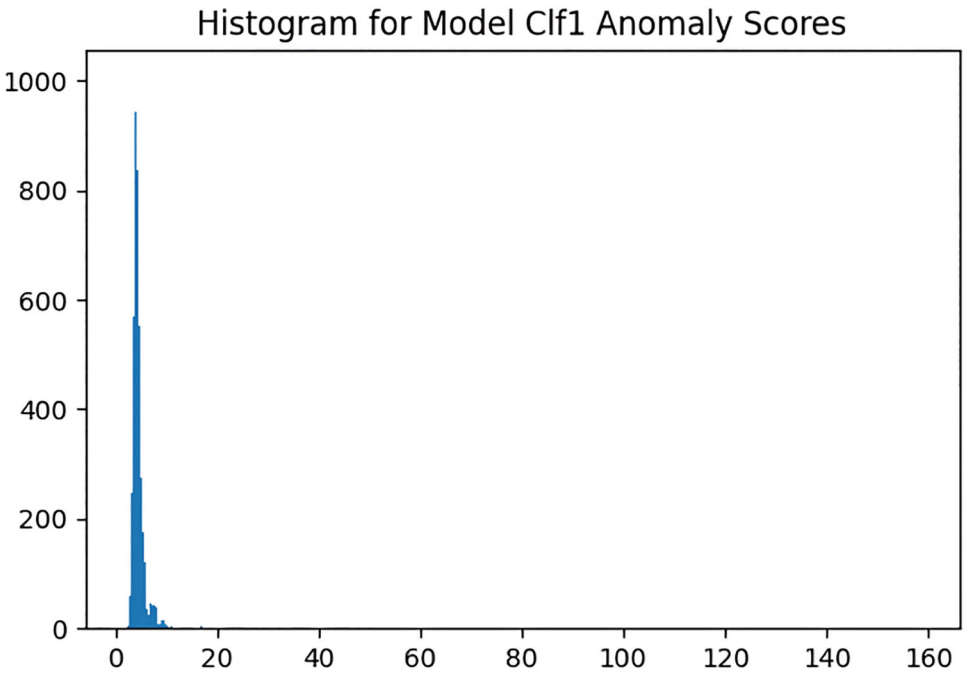


Figure 6. Histogram of anomaly scores for trial 1.

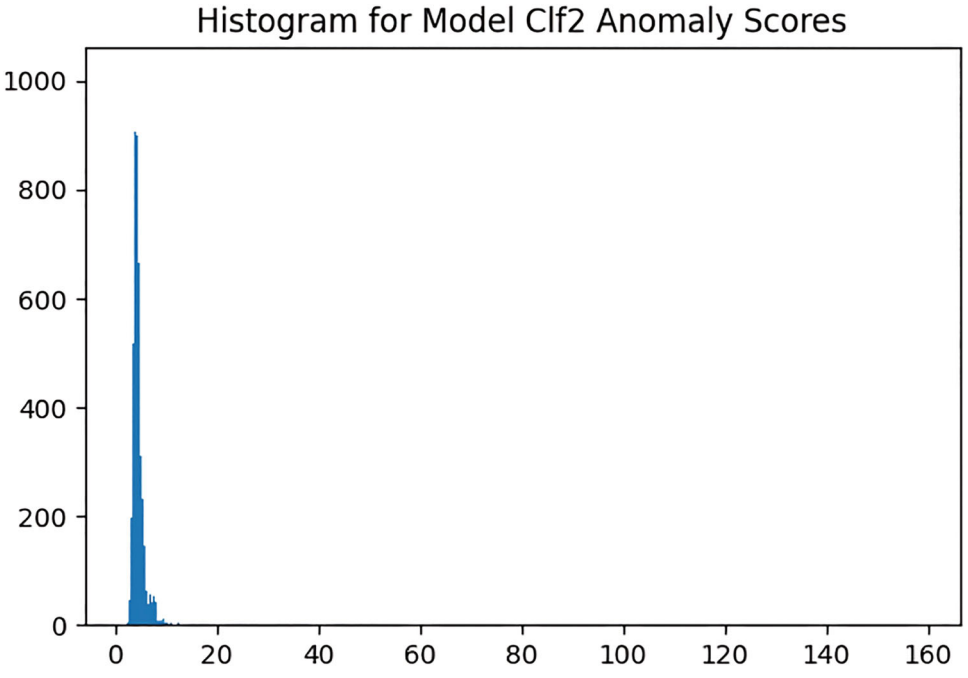


Figure 7. Histogram of anomaly scores for trial 2.

and below the mean, respectively. The margin of error tells us over what range we expect the true population parameter of interest to exist based on the sample data that was collected. This is determined through the use of z-score and standard error of the mean. A confidence interval provides a range of values within which one can expect the true population parameter of interest to exist with a specified level of confidence. Here, 0.95 was taken as the confidence interval value.

The histograms of anomaly scores are plotted for both the models and are depicted in Figures 6 and 7. The threshold for the same are calculated using the derived z-score and margin_of_error values.

5. Implementation

5.1. Healthcare dataset source

For our research, we used data published by the Centers for Medicare and Medicaid Services (CMS). The Medicare Physician and Other Practitioners by Provider and Service dataset provides information on use, payments, and submitted charges organized by National Provider Identifier (NPI), Healthcare Common Procedure Coding System (HCPCS) code, and place of service. The dataset contains approximately 10 million records, capturing a wide range of service details for different provider types. These provider types are

associated with each NPI and include physicians, nurse practitioners, physician assistants, clinical social workers, and other healthcare practitioners. This data is unlabeled. Thus, we employed unsupervised anomaly detection to get the Anomaly label for these records, which gives us the fraudulent records. Further, a suitable classifier is used to train the obtained labeled data and evaluate its fraud detection performance.

5.2. Tools used

The paper presents the development of the models of anomaly detection using Python Outlier Detection (PyOD). PyOD is a Python package built to detect outliers and anomalies in data in an easy-to-use way. It's a large collection of anomaly detection algorithms ranging from classical statistical methods to contemporary machine learning approaches. Since PyOD is user-friendly and adaptable, it can be put to various uses in a lot of different areas and provides a range of outlier detection techniques. Any user is able to choose the best algorithm for a particular use case.

The study also utilized Pandas, NumPy, Scikit-learn and Seaborn in working with data. Pandas is a fundamental library of Python for data manipulation and analysis. Since it has intuitive syntax and wide function availability, Pandas makes users perform common tasks of data such as cleaning, transformation and exploration efficiently and fast. It provides data missing handling, grouping, merging, and reshaping datasets and hence is an indispensable library for data preprocessing and preparation. Besides, Pandas has very tight integration with other Python libraries, which allows easy data exchange and interoperability.

A core library for numerical computing in Python, NumPy supports matrices, multi-dimensional arrays and a vast range of mathematical functions. NumPy effectively allows the implementation of array-based computations and helps users in scientific computing and high-performance numerical processing jobs. Because of the flexibility of its array objects, users can easily perform element-wise calculations, statistical calculations, and linear algebra operations. Further, NumPy easily interfaces with other scientific tools such as SciPy and Matplotlib that facilitate complex algorithms and visualizations.

Scikit-learn is probably the best machine learning library that provides the complete toolkit for building and deploying a machine learning model. With scikit-learn, users can build complex models in minimum time. Since its library is extensive and it has good and well-documented functionality, strong speed and consistent API, it is preferred in different sectors for machine learning jobs. Scikit-learn makes every step from choosing a model to model evaluation, to hyperparameter tuning and pipeline creation simple, thus making prototyping much faster and the experiment even quicker.

Seaborn is built on top of Matplotlib and is one of the most powerful libraries for data visualization because of its high-level interface for drawing attractive and informative statistical graphics. Seaborn offers simplification of creating complex visualizations through a variety of plotting functions with an easy syntax so users can communicate findings effortlessly. Seaborn provides a flexible toolkit for data visualization distribution, correlation and patterns from very basic plots like scatter plots and histograms, up to more complicated ones like violin plots and pair plots. Further, it is a very useful library in exploratory data analysis and presentation because of its smooth interaction with Pandas data structures and the customization of plot aesthetics.

The development environment used for running experiments and training of the machine learning models in this work is Google Colab, a web-based Jupyter Notebook environment. One important advantage of using Google Colab is that one can exploit the availability of T4 GPUs for efficient acceleration in training and experimenting with the models. Stronger computational powers offered through the use of the T4 GPU accelerators allow for quicker convergence, and thus better utilization of resources, especially during model training while focusing on deep learning model types like autoencoders. The project enjoys much shorter times for training, while it is more productive, since the usage of T4 GPUs in Google Colab allows faster iteration to improve anomaly detection algorithms. Moreover, the development process might be further empowered with smooth integration provided by Google Colab with famous machine learning tools TensorFlow and PyTorch.

Table 1. Decision tree classifier scores.

Model	Accuracy	Recall	Precision	F1-score
Decision Tree Classifier	0.9986	0.9750	1.0000	0.9857

Table 2. Parameter values obtained in autoencoder.

Values taken	Trial 1	Trial 2
Hidden neurons	[15, 10, 6, 2, 2, 6, 10, 15]	[10, 6, 2, 2, 6, 10]
Number of epochs	25	25
mean_score	4.57	4.59
std_score	1.85	1.84
z_score	1.96	1.96
Margin of error	0.011	0.011
Threshold	4.58	4.60

6. Results and discussions

The Isolation Forest model could detect all of the existing fraudulent records i.e. five percent of the total data records. It had the maximum accuracy in anomaly detection. The 3D TSNE Plot for Outliers obtained using Isolation Forest is depicted in Figure 8.

The Local Outlier Factor model could detect 88% percent of the existing fraudulent data records. The 3D TSNE Plot for Outliers obtained using Local Outlier Factor is depicted in Figure 9.

The K-Nearest Neighbours model detected 70% percent of the existing fraudulent data records. The 3D TSNE Plot for Outliers obtained using K-Nearest Neighbours is depicted in Figure 10.

The Decision Tree Classifier gave best results on the labeled healthcare data, which was obtained from the results of Isolation Forest model. After training the classifier on 70% of the data, it could accurately classify all the fraudulent records in the remaining 30% data, which is the test data. Table 1 shows the Decision Tree Classifier scores. Figure 11 shows the confusion matrix for the same.

Significant insights into the autoencoder model's performance for anomaly detection can be found from Table 2's results, which highlight the impact of the number of epochs and differences in the hidden neuron architecture on the model's performance.

Where model complexity becomes the trade-off for performance, the variance in the architecture of hidden neurons between trials can be clearly gauged. Trial 2 has a simpler architecture with fewer layers and neurons, which likely contributes to faster convergence and reduced computational cost. This may be favored in real-world scenarios where computational resources are limited. In contrast, Trial 1 uses a deep architecture with more hidden layers and neurons, thus allowing better representation of features and capturing more complex patterns in the data. This holds special relevance in healthcare fraud detection since, in most scenarios, anomalies tend to be complicated and usually subtle. Despite these differences, both trials yield similar mean scores, standard deviations, z-scores, margins of error and thresholds. In other words, they are pretty consistent in terms of

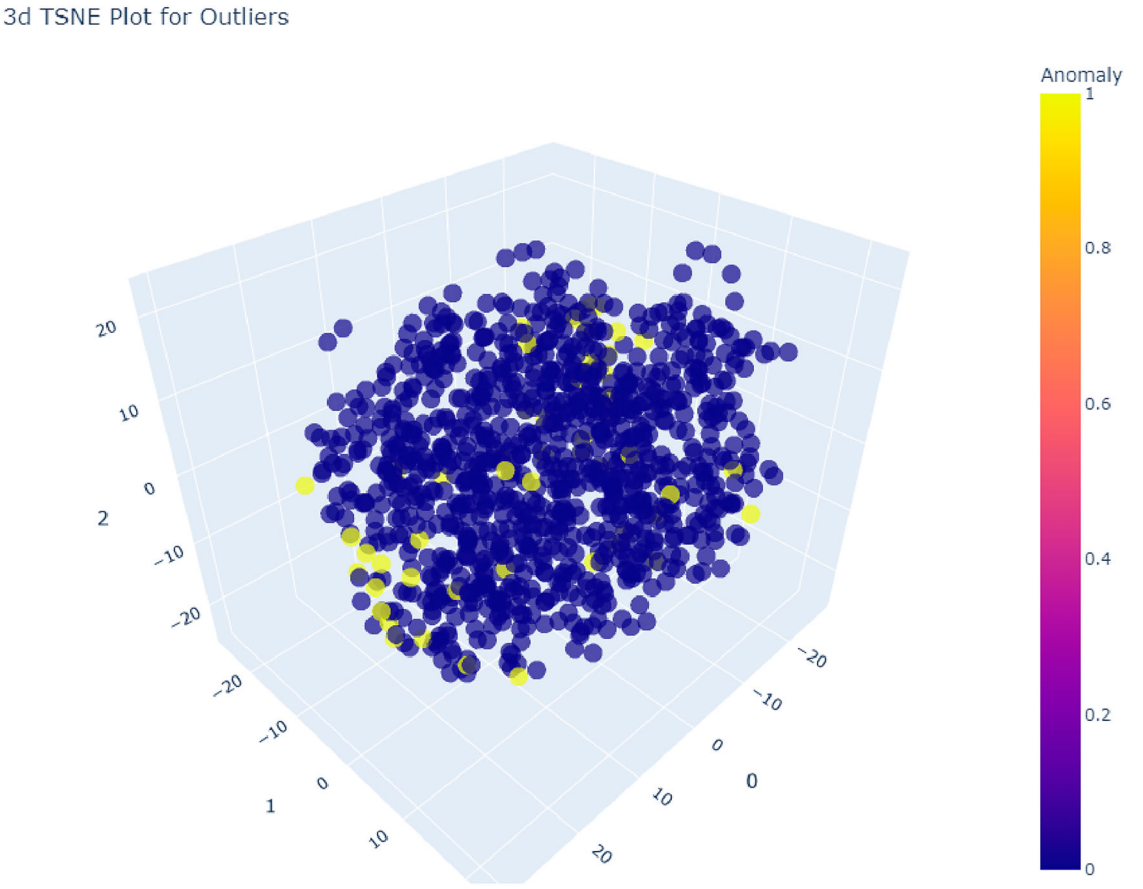


Figure 8. Isolation forest- 3D TSNE plot for outliers.

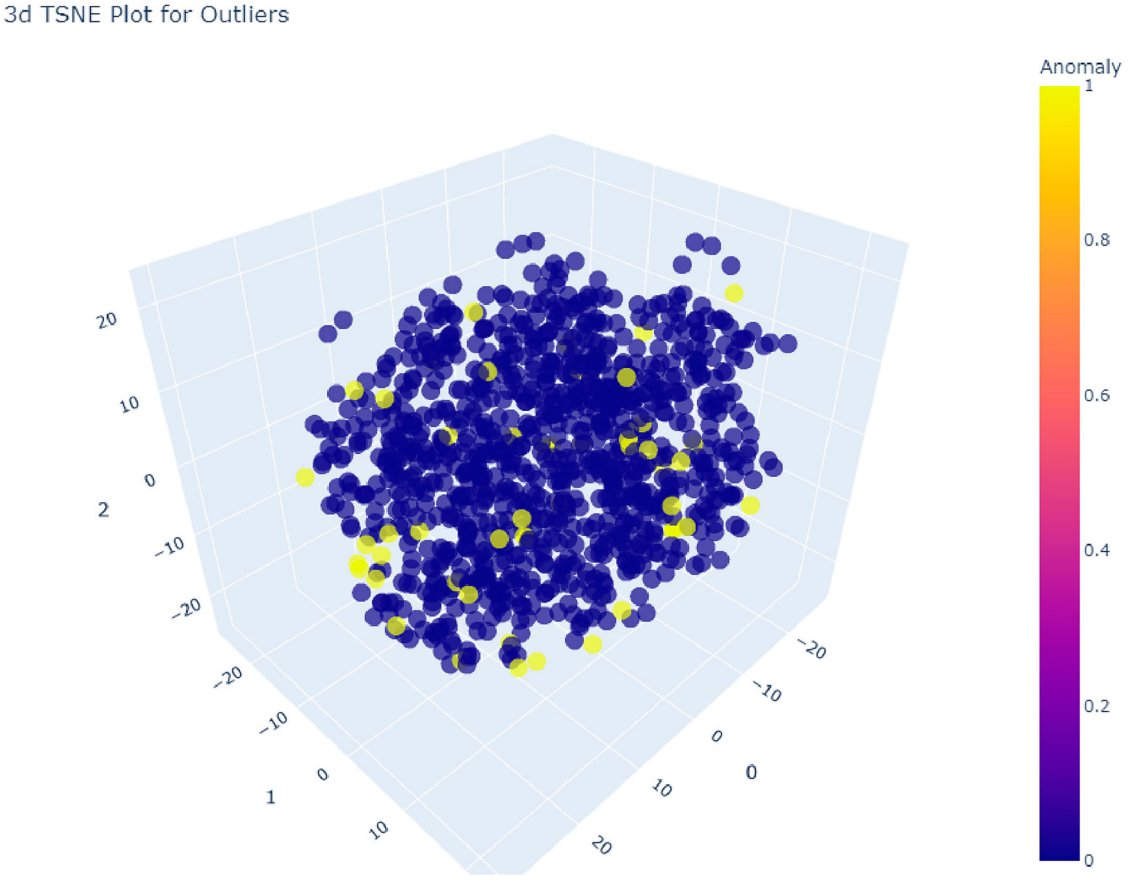


Figure 9. Local outlier factor- 3D TSNE plot for outliers.

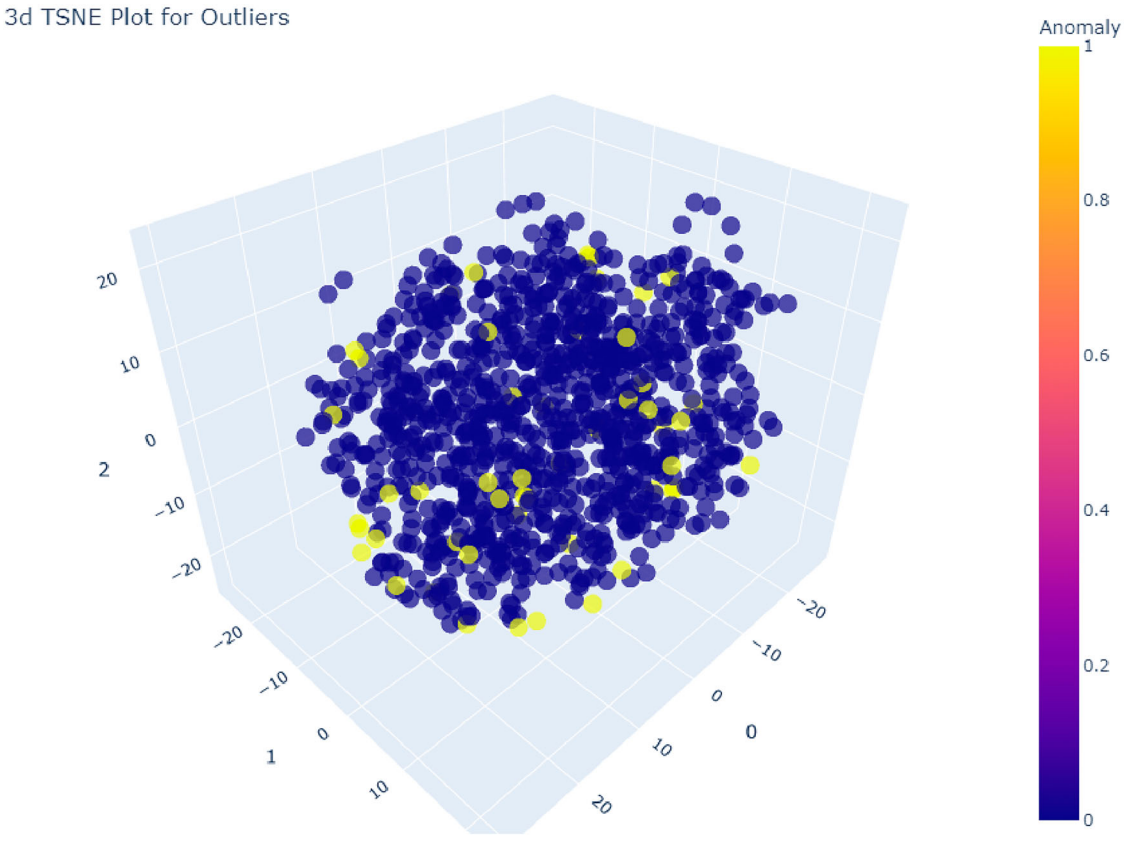


Figure 10. K-nearest neighbors- 3D TSNE plot for outliers.

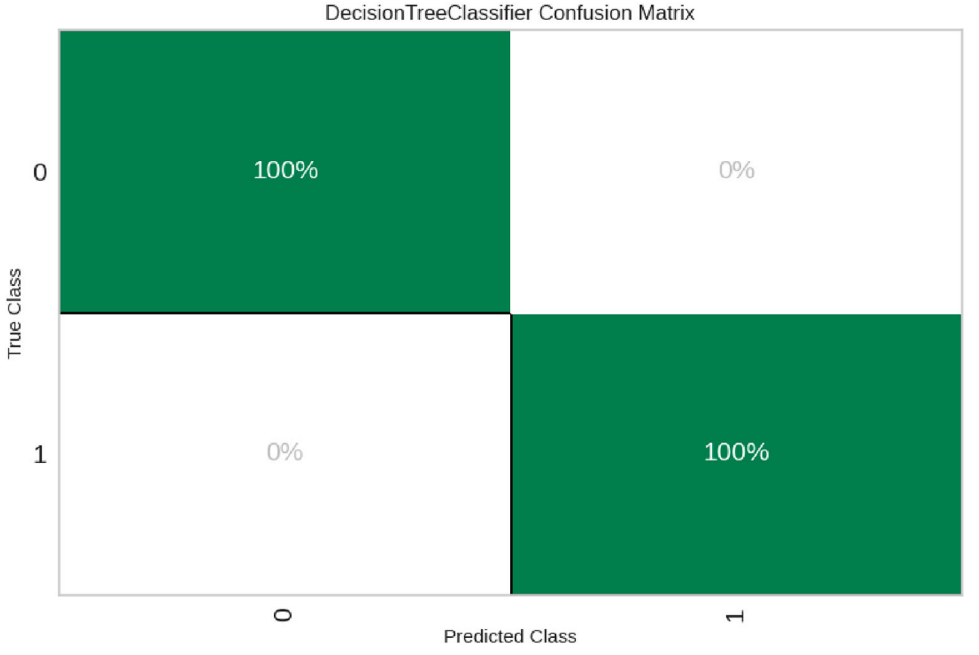


Figure 11. Decision tree classifier- confusion matrix.

finding anomalies. The *z*-score and margin of error provide a measure of confidence in the results, with the threshold acting as a critical level beyond which a sample is classified as an anomaly.

The main reason for Isolation Forest outperforming the other models in anomaly detection is most likely because it does an effective job in the isolation of anomalies within high-dimensional spaces. This also makes it particularly suitable for healthcare data, which usually comprises of several dimensions with complicated relationships. The Isolation Forest technique for segregating anomalies which is based on the assumption that anomalies are few and different, might hold an edge in the detection of unusual patterns in large and diverse datasets. This explains why it performed better than the other models. The specific characteristics of the healthcare data, such as its high dimensionality and the presence of subtle anomalies

likely contributed to this outcome. These tend to indicate that for balancing performance and resource utilization, more complex models or algorithms, such as the Isolation Forest, may be required to represent the complex nature of healthcare fraud.

7. Conclusion and future scope

7.1. Conclusion

It can be concluded that out of the machine learning anomaly detection models taken, Isolation forest performs the best in terms of fraud detection when applied on the original healthcare data. Decision Tree Classifier works best for classifying labeled healthcare data as normal or fraudulent. Complementing these techniques, the autoencoder model provides a deep learning-based solution that has the ability to learn complex features and nuances in the healthcare data. However, it should be noted that the findings of this study are conducted based on just one dataset, which could possibly limit the generalizability of the results to other healthcare data. Moreover, future studies could apply these models to more diverse datasets with the aim of analyzing their scalability and adaptability. Moreover, further investigation and optimization of the parameters of autoencoder could promise even better performance in anomaly detection, which would contribute to a robust anomaly detection framework for healthcare applications.

7.2. Future scope

While this study adopts appropriate machine learning and deep learning techniques that enhance fraud detection methods in terms of better performance and flexibility, a few limitations must be declared. Due to the use of only one dataset, there cannot be a generalization of the results for various healthcare contexts or other data sources. The selected models are not extensively tested for their scalability for larger and various datasets, which is a very crucial feature in real-world applications. In the future, therefore, studies should be directed towards increasing the scope of datasets by involving different types of healthcare data and exploring other deep learning methods which may turn out to be superior for specific applications. Other future work will involve the actual implementation of these models in real life, which may bring about possible challenges in system integration, data privacy, and real-time processing capability. The ultimate goal should be to strengthen the resilience of health systems against emerging fraud schemes through proactive solutions that prevent fraud, waste and abuse.

Healthcare fraud seriously threatens the integrity and sustainability of healthcare programs worldwide. The fraudulent activities with respect to fake insurance claims, billing scams, and identity theft in the healthcare industry are hugely costly, running into billions of dollars each year. These unethical activities put treatments and public confidence in the healthcare system at risk, besides diverting very important funds away from healthcare initiatives. Healthcare fraud requires a multilayered approach involving regulatory monitoring, domain expertize, and advanced machine learning. By leveraging machine learning algorithms with deep-diving capabilities into the volumes of healthcare data, this research has tried to find the anomalies that break away from regular patterns. These use various features such as provider information, billing codes, medical procedure details, and patient demographics to identify those transactions that are possibly fraudulent. The insights drawn from this study are useful in the construction of robust fraud detection systems targeting patient-centered healthcare data. Policymakers and service providers can use these insights in devising effective fraud prevention strategies and further enhance the overall security of healthcare programs. However, it will be important in future research to continue exploring novel fraud detection techniques and adapt these systems to new fraudulent strategies to maintain long-term efficiency.

Disclosure statement

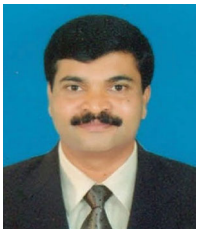
No potential conflict of interest was reported by the author(s).

Data availability statement

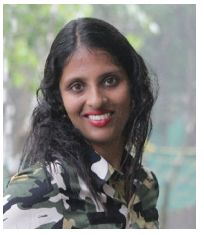
The data that support the findings of this study are available at https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners/medicare-physician-other-practitioners-by-provider-and-service.

Notes on contributors

Maithri Bairy is currently pursuing her M.Tech in Computer Science and Engineering from Manipal Institute of Technology, Manipal. She holds a B.Tech in Information Technology and her areas of interest include Artificial Intelligence and Data Science.



Dr. Balachandra Muniyal received his B.E degree in Computer Science and Engineering from Mysore University and M.Tech, and Ph.D in Computer Science and Engineering from Manipal Academy of Higher Education, Manipal, India. His research area is Cyber Security. He has more than 90 publications in national and international conferences/journals. Currently, he is working as a Professor in the Dept. of Information & Communication Technology, Manipal Institute of Technology, Manipal. He is also coordinating the Centre of Excellence for Cybersecurity, MAHE, Manipal. He was the Head of the Department from 2017 to 2020. He has 30 years of teaching experience in various Institutes. Under his supervision, five research students completed their Ph.D and currently, he is guiding eight students.



Dr. Nisha P. Shetty is a dedicated academician currently serving as an Assistant Professor (Senior Scale) in the Department of Information and Communication Technology at Manipal Institute of Technology (MIT), a part of the Manipal Academy of Higher Education (MAHE). With a solid educational background, Dr. Shetty earned her Ph.D in Data Privacy and Security from MIT, MAHE in 2023, an M.Tech in Computer Science and Engineering from Visvesvaraya Technological University (VTU) in 2015, and a B.E in Computer Science and Engineering from VTU in 2013. Additionally, Dr. Shetty serves as a faculty advisor for Project Cryptonite, one of the top-performing student projects at MIT, which has been winning accolades nationally and internationally. Her research interests include data privacy and security, with a focus on privacy-preserving frameworks in social networks, deep learning for medical diagnostics, and enhanced security measures for online data.

ORCID

Maithri Bairy http://orcid.org/0009-0003-0208-4455
Balachandra Muniyal http://orcid.org/0000-0002-4839-0082

References

[1] Branting LK, Reeder F, Gold J. Graph analytics for healthcare fraud risk estimation. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM); San Francisco, CA, USA; 2016. p. 845–851.
[2] Thaifur AY, Maidin MA, Sidin AI, et al. How to detect healthcare fraud? "a systematic review". Gac Sanit. 2021;35:S441–S449. doi: 10.1016/j.gaceta.2021.07.022the 3rd International Nursing and Health Sciences Students and Health Care Professionals Conference (INHSP).
[3] Liu J, Bier E, Wilson A, et al. Graph analysis for detecting fraud, waste, and abuse in health-care data. Ai Mag. 2016;37:33–46.

[4] Agrawal N, Panigrahi S. A comparative analysis of fraud detection in healthcare using data balancing & machine learning techniques. In: 2023 International Conference on Communication, Circuits, and Systems (IC3S); Bhubaneswar, India; 2023. p. 1–4.
[5] Dornadula VN, Geetha S. Credit card fraud detection using machine learning algorithms. Procedia Comput Sci. 2019;165:631–641. doi: 10.1016/j.procs.2020.01.0572nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP – TIV INNOVATION, 2019 November 11–12, 2019
[6] Mehbodniya A, Alam I, Pande S, et al. Financial fraud detection in healthcare using machine learning and deep learning techniques. Secur Commun Netw. 2021;2021:1–8. doi: 10.1155/2021/9293877
[7] Oza-aditya A. Fraud detection using machine learning. 2018. [Online]. Available from: https://api.semanticscholar.org/CorpusID:202618514.
[8] Raghavan P, Gayar NE. Fraud detection using machine learning and deep learning. In: 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE); Dubai, United Arab Emirates; 2019. p. 334–339.
[9] Raman R, Tiwari M, Buddhi D, et al. Cyber security fraud detection using machine learning approach. In: 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE); Greater Noida, India; 2023. p. 1037–1042.
[10] Yoo Y, Shin J, Kyeong S. Medicare fraud detection using graph analysis: a comparative study of machine learning and graph neural networks. IEEE Access. 2023;11:88278–88294. doi: 10.1109/ACCESS.2023.3305962
[11] Kemp J, Barker C, Good N, et al. Developing an anomaly detection framework for medicare claims. In: Proceedings of the 2023 Australasian Computer Science Week, ser. ACSW '23. New York, NY, USA: Association for Computing Machinery; 2023, p. 234–237. [Online]. doi: 10.1145/3579375.3579410
[12] Waghade SS. A comprehensive study of healthcare fraud detection based on machine learning. 2018. [Online]. Available from: https://api.semanticscholar.org/CorpusID:201048558.
[13] Bauder R, Khoshgoftaar T. Medicare fraud detection using random forest with class imbalanced big data. In: 2018 IEEE International Conference on Information Reuse and Integration (IRI); Salt Lake City, UT, USA; 2018. p. 80–87.
[14] Reychav I, McHaney R, Babbar S, et al. Graph network techniques to model and analyze emergency department patient flow. Mathematics. 2022;10(9):1526. doi: 10.3390/math10091526
[15] Pourhabibi T, Ong K-L, Kam BH, et al. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. Decis Support Syst. 2020;133:113303. doi: 10.1016/j.dss.2020.113303
[16] Ali N, Praveen S. Outlier detection on clinical data using deep learning algorithms. Int Res J Eng Technol (IRJET). 2021;8(6):1901–1910.
[17] Munir M, Chattha MA, Dengel A, et al. A comparative analysis of traditional and deep learning-based anomaly detection methods for streaming data. In: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA); Boca Raton, FL, USA; 2019. p. 561–566.
[18] Bowie M, Begoli E, Park B, et al. Towards an LSTM-based approach for detection of temporally anomalous data in medical datasets. 2017.
[19] R VK, Shaik S, Rao B, et al. Machine learning based outlier detection for medical data. Indones J Electr Eng Comput Sci. 2021;24:564.
[20] Torabi H, Mirtaheri S, Greco S. Practical autoencoder based anomaly detection by using vector reconstruction error. Cybersecurity. 2023;6(1):1. doi: 10.1186/s42400-022-00134-9
[21] Sattarov T, Herurkar D, Hees J. Explaining anomalies using denoising autoencoders for financial tabular data. 2022.
[22] Berahmand K, Daneshfar F, Salehi E, et al. Autoencoders and their applications in machine learning: a survey. Artif Intell Rev. 2024;57(2):28. doi: 10.1007/s10462-023-10662-6
[23] Kopčan J, Skvarek O, Klimo M. Anomaly detection using autoencoders and deep convolution generative adversarial networks. Transp Res Procedia. 2021;55:1296–1303. doi: 10.1016/j.trpro.2021.07.113
[24] Gonzalez D, Patricio MA, Berlanga A, et al. Variational autoencoders for anomaly detection in the behaviour of the elderly using electricity consumption data. Expert Syst. 2021;39(4):e12744. doi: 10.1111/exsy.v39.4
[25] Alhassan Z, Budgen D, Alshammari R, et al. Stacked denoising autoencoders for mortality risk prediction using imbalanced clinical data. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA); Orlando, FL, USA; 2018. p. 541–546.
[26] Sakurada M, Yairi T. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In: Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis, ser. MLSDA'14. New York, NY, USA: Association for Computing Machinery; 2014. p. 4–11. [Online]. doi: 10.1145/2689746.2689747
[27] Senaratne A, Christen P, Williams GJ, et al. Rule-based knowledge discovery via anomaly detection in tabular data. In: Make. 2023. [Online]. Available from: https://api.semanticscholar.org/CorpusID:260356626.
[28] Nawaz A, Khan SS, Ahmad A. Ensemble of autoencoders for anomaly detection in biomedical data: a narrative review. IEEE Access. 2024;12:17273–17289. doi: 10.1109/ACCESS.2024.3360691
[29] Mues K, Liede A, Liu J, et al. Use of the medicare database in epidemiologic and health services research: a valuable source of real-world evidence on the older and disabled populations in the us. Clin Epidemiol. 2017;9:267–277. publisher Copyright: © 2017 Mues et al. doi: 10.2147/CLEP
[30] Samariya D, Ma J, Aryal S, et al. Detection and explanation of anomalies in healthcare data. Health Inf Sci Syst. 2023;11(1):20. doi: 10.1007/s13755-023-00221-2
[31] Kumaraswamy N, Markey MK, Ekin T, et al. Healthcare fraud data mining methods: a look back and look ahead. PubMed. 2022;19(1):1i.
[32] Xiaohan Yu KD, Wang H, Chen C. A novel ldvp-based anomaly detection method for data streams. Int J Comput Appl. 2024;46(6):381–389.