HEALTH INSURANCE FRAUD DETECTION USING MACHINE LEARNING AND DEEP LEARNING TECHNIQUES

A PROJECT REPORT

Submitted in partial fulfilment of the requirements for the award of the internship

By

Pentakota Yasoda (21VV1A1250)

Pichika Parimala Durga Srivalli (21VV1A1251)

(Students of JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY – GURAJADA VIZIANAGARAM (A))

Supervisor:

Dr. Mohammad Farukh Hashmi

Assistant Professor

Department of Electronics and Communication Engineering

Department of Electronics and Communication Engineering

National Institute of Technology, Warangal

2023 - 2024

ABSTRACT

Fraud detection is an important area of research in the healthcare system due to its financial consequences, arising mainly from costs, losses, and risks. Most businesses use fraud detection algorithms based on machine learning and/or deep learning to lessen this. Healthcare systems function better when fraud detection models are effective. Principal obstacles in developing a successful fraud detection model including:

- **Data imbalance:** A skewed proportion of less fraudulent cases relative to cases that are not fraudulent
- **Model selection for classification:** using suitable deep learning or machine learning models to distinguish between fraud and non-fraud cases

In this work, we have used different data-imbalance techniques and classification models to meet these challenges; we have also used variants of neural network models. There were various models that were tested as part of this study. These models' performance was evaluated using a range of criteria, including F1-score, recall, accuracy, and precision. In this study, it was found that a neural network model trained on under sampled data outperformed the other models.

Keywords: machine learning, deep learning, healthcare, fraud detection, neural networks, convolutional neural networks (CNN), recurrent neural networks (RNN), medical datasets, data imbalance, classification models.

I. INTRODUCTION

Healthcare sector is one of the prominent sectors in which a lot of data can be collected not only in terms of health but also in terms of finances [1]. Major frauds happen in healthcare sector. It is one of the important service providers that improves people lives. Health insurance becomes the only means of receiving high-quality care in the event of an accident or serious sickness when the cost of healthcare services rises.

As health insurance will reduce the costs and provides financial and economic stability for an individual. One of the main tasks of health care insurance providers is to monitor and manage the data and to provide support to customers. Regulations and business secrecy prevent insurance firms from disclosing patient data, but since data are not synchronised and integrated among insurance providers, there has been a rise in healthcare fraud.

Often false information is provided to health insurance companies in order to make them pay for some false claims to the policy holders. A policyholder may also be able to get benefits from more than one insurance company. The **National Health Care Anti-Fraud Association (NHCAA)** estimates that there is a yearly financial loss of billions of dollars [15]. Developing a system to safely handle and keep an eye on insurance operations through the integration of data from all insurance providers is essential to preventing health insurance fraud.

According to the 2019 report of the NHCAA on healthcare fraud detection, the total losses in 2018 was USD 679.18 million, which is expected to reach USD 2.54 billion by 2024. In India, according to the report of business today [2] in 2019, health insurance fraud cost around 45,000 crores.

MOTIVATION

The desire to overcome the shortcomings of conventional approaches and improve the efficiency of fraud prevention tactics is the driving force behind the application of machine learning (ML) and deep learning (DL) techniques in fraud detection. The primary driving force behind the use of deep learning and machine learning approaches in fraud detection is their capacity to offer sophisticated, flexible, and scalable solutions that improve security, effectiveness, and precision in detecting and stopping fraudulent activity. Among the reasons for motivation are:

- **Increased Accuracy:** By continuously learning from fresh data, ML and DL models can become more and more adept at differentiating between authentic and fraudulent activity. As a result, there are fewer false positives and more accuracy, which cut down on pointless interventions.
- **Enhanced Security:** ML and DL models are able to spot hidden and subtle trends that point to fraud that traditional systems or human analysts could miss. As a result, security is enhanced and there is a greater chance of discovering fraudulent activity early on.
- **Growing Complexity of Fraud:** In order to avoid discovery, fraudsters are always coming up with increasingly advanced methods. Conventional rule-based systems find it difficult to stay up to date with these new techniques. With their ability to adjust to novel patterns and strategies, ML and DL offer a strong defensive against intricate fraud schemes.

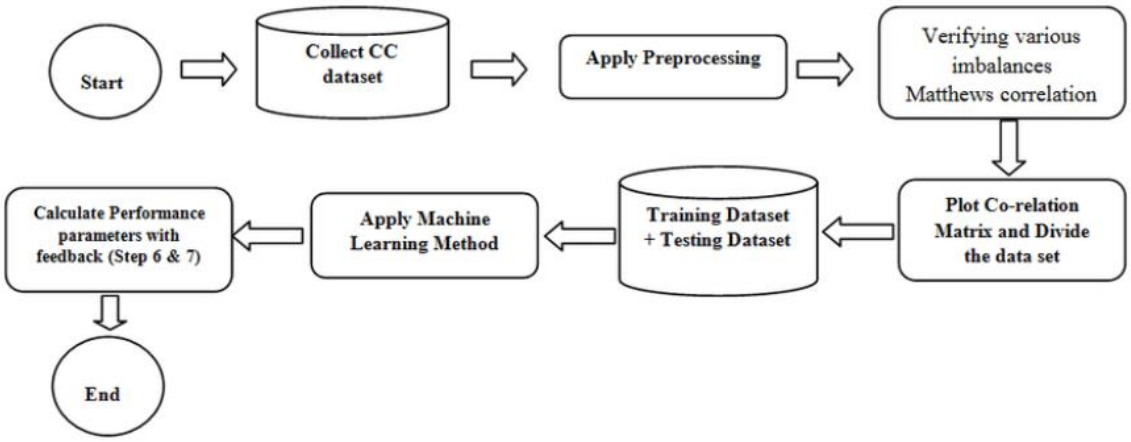

Figure 1 – Procedure for fraud detection using ML

A. CONTRIBUTION

Three major categories can be used to categorise healthcare fraud: insurer fraud, customer fraud, and provider fraud [3]. Healthcare fraud perpetrated by providers can come from both individuals and organisations, such as hospitals or physicians.

Sometimes provider fraud may also involve other service providers or individuals (e.g., patients). Customer fraud may be committed when the insured/ consumer knowingly misrepresents the facts to get additional benefits. They might collaborate with healthcare practitioners (such as physicians) in a union. The focus of this paper is on customer-level fraud by observing the claim records as collected by the healthcare systems.

In this paper, various machine learning and deep learning approaches are used for detecting frauds and different algorithms such as Naïve Bayes, Regression, K-Nearest Neighbor (KNN), Ada Boost, Decision Tree, Support Vector Machine (SVM), and Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) are skewed for training to detect the frauds. The goal of this work is to show how different deep learning and machine learning models may be used to healthcare systems to effectively detect fraud.

This is aimed to identify the best model that can be used for identifying fraudulent claims. We have presented the performance of various models using standard performance metrics and the best performance among them.

In this paper, we first provide a brief review of the literature in the area of healthcare fraud. Then we describe the two-phase methodology used to arrive at the fraud detection model.

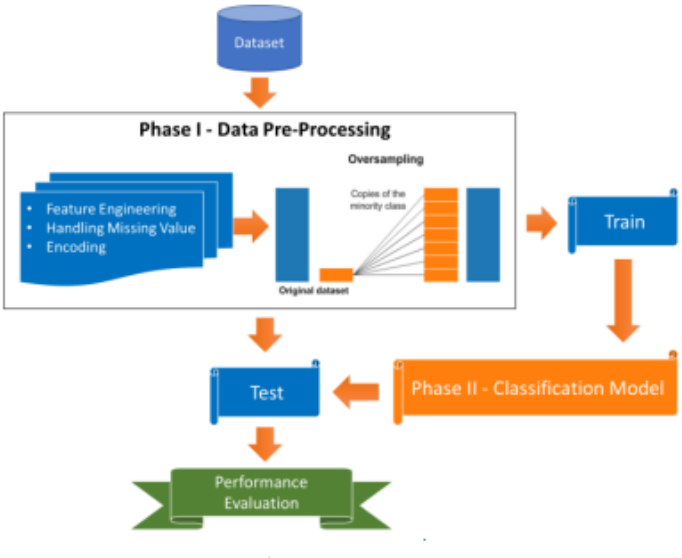

Figure 2 – Proposed method

In the data section, we explain in detail the dataset used in this work and provide descriptive statistics of the data. The outcomes of the various model implementations used in this study are then included in the results section. Lastly, we have the conclusion, which is followed by more research in this study.

II. PROBLEM STATEMENT

The healthcare sector is always evolving with respect to technology advancements and taking new forms. [4]. It is necessary to maintain and monitor the patient's record without any ambiguity. User access to high-quality healthcare services is required. Because of the growing technology, it is necessary to build a system in which the data is secured and maintained accurately. Due to the lack of traceability in the data transaction and the records, there have been several problems in the healthcare system.

The healthcare insurance companies are using a method to keep track of the healthcare records and insurance approved details of the customers. This makes a way for many frauds to occur in the health insurance domain. Hence, it is very necessary to build a secure system and keep track of all the claims insurance from a user so that the fraud can be detected. Accordingly, the number of frauds occurring in healthcare can be reduced and more appropriate quality services can be provided to the users.

This requires a need to build a system, and make use of technology to store and trace the data. The solution is to make use of the machine learning and deep learning techniques, which helps in data integrity and security. Therefore, it helps us to overcome many of the challenges that exist in the healthcare industry including data integrity, data security, and fraud detection.

III. RELATED WORK

This section reviews different fraud detection technologies with various models and different machine learning approaches. Today, a growing number of healthcare insurers are using the latest machine learning and data mining tools to build fraud detection models [5]. Effective fraud detection models only get stronger at finding hidden patterns in data that are otherwise invisible as they have more data for training over time. Efforts are being made by various researches in this domain and rule-engines are used in various healthcare systems, which has a set of pre-defined rule. They are implemented to identify errors, incomplete data, duplicate claims, ineligible claims, suspicious claims, etc.

These systems may not have the capability to model fraudulent behavior. More sophisticated fraud detection models are based on data mining and machine learning [5]. The proposed method is a methodology to detect fraud and is tested on real-world insurance data. Various models are used to identify fraudulent patterns in healthcare claims. Python is used to carry out the task.

The goal of this research is to concentrate on machine learning and deep learning methodologies in which various methodologies are utilized. The methodologies' accuracy, precision, recall, and F1-score are utilized to compare their outcomes. The ROC curve is produced by means of the confusion matrix. These approaches' performance metrics – accuracy, precision, recall, and F1-score are compared. The best methodology for detecting fraud is generally considered to be the one having the best performance metrics.

Table: Limitations of Machine Learning Techniques

Models	Strength	Limitations
Bayesian	Provide better results in problems of binary classification and suitable for analyzing the real-time data	Requires better detection related to abnormal behavior of fraud cases
Neural Network	Suitable for problems related to binary classification, mostly used for detecting the fraud	Requires huge computation, can be denied for real-time operation
Decision Tree	Implementation is more straightforward with low power of computation and suitable for analyzing the real-time data	Overfitting may rise if the information of underlying domain does not set in the training data
Linear Regression	When dependent and independent variables have an almost linear relationship, it generates an optimal result	Sensitive for the outliers and numeric value limitation
Logistic Regression	Implementation is easy and fraud detection is based on historical data	Performance of classification is lacking when compared with methods of data mining
Support Vector Machine	The non-linear problem of classification is solved with low power of computation and suitable for analyzing real-time data	Input data transformation results in difficulties while processing the data

Table 1 – Limitations and Strengths of various machine learning models

It is believed that around 16% od the United States GDP is spent on healthcare, out of which some numbers of resources are wasted because of medical errors, fraud and abuse, payment for services which are not delivered [6]. It is estimated that there is around 3% to 10% of health insurance claims are a fraud and this condition is similar in most of the countries around the world [7]. According to the NHCAA**,** there will be around tens of billions of dollars of loss caused by health insurance fraud cases each year. The NHCAA has provided a list of fraud cases which includes false patient diagnosis, and medical histories, theft of patients' health insurance benefits, claiming insurance from multiple insurance companies, etc. [8].

Architecture model of related work.[5]

Related work	Model	Security
HDEHR	DE, P2P	—
m-Health	DE	—
uPHR	DE	—
CF	CS, DO	CIA, HIPAA
HealthVault	CS	Authentication
healthTicket	CS	CP-ABE
DEPR	DC	—
My HealtheVet	DE	Security policies
SNOW	DC	Privacy policies

HDEHR: hierarchical distributed electronic health record; uPHR: ubiquitous personal health record; CIA: Confidentiality, Integrity and Availability; DEHR: Doversity and Equity in Health Reform; CS: client–server; DO: distributed object; DC: distributed component; HIPAA: Health Insurance Portability and Accountability; CP-ABE: Ciphertext-policy Attribute-based Encryption; P2P: peer-to-peer.

Table 2 – Architecture model of related work

In this figure, the model concentrates on health data of all patients in single and multiple servers. Many approaches are available to detect healthcare fraud claims in machine learning and deep learning fields. One of the methods available to find fraud detection in healthcare insurance claim by data-driven method on the research, fraud cases and literature review [9].

In this work. We have used various machine learning and deep learning models to build an efficient fraud detection model for healthcare services.

IV. EXPERIMENTAL SETUP AND METHODS

This section describes how to use a dataset and various machine learning and deep learning classifiers, including the sequential model, KNN, Decision Tree, Naive Bayes, and Linear and Multiple Regression, in the experiment. Before creating the classifier, each of these algorithms completes a variety of tasks, including gathering data, preparing it, evaluating it, training it using various classifiers, and finally testing it.

The preprocessing step involves transforming all of the data into a format that can be used. Two distinct data distribution sets were used for the hybrid under sampling (negative class) and oversampling (positive class) approaches. Pre-processed data is given into the classifier algorithm during the training phase. Subsequently, the testing data is assessed to determine the accuracy in identifying credit card fraud.

In the end, accuracy and optimal performance are used to compare and assess each model. To determine whether model performs better when evaluated in a real-time setting, a subset of fraud transactions is combined with the legal ratio.

A. SEQUENTIAL MODEL

In this project, Neural networks were essential in detecting fraudulent activity in healthcare data for Fraud Detection Using ML and DL Techniques. Because neural networks can handle huge and complicated datasets and can pick up on subtle patterns that typical machine learning (ML) models would overlook, they are used as a subset of deep learning (DL) approaches. Neural Network Architecture:

- **Input Layer:** Pre-processed features from medical records, including patient demographics, treatment information, billing details, and past claim data, were fed into the neural network's initial input layer.
- **Hidden Layer:** To capture intricate relationships within the data, many hidden layers were used. Every neuron in a hidden layer applied a nonlinear activation function (like RELU) to induce nonlinearity and help the network learn more intricate patterns. Through trial and error, the depth and width of these layers were adjusted to balance computational efficiency and performance.
- **Output layer:** The output layer, which produced a probability score indicating the possibility of a claim being fake, was the last layer. It used a sigmoid activation function. It was simple to use thresholding to categorize claims as fake or not thanks to this probabilistic output.

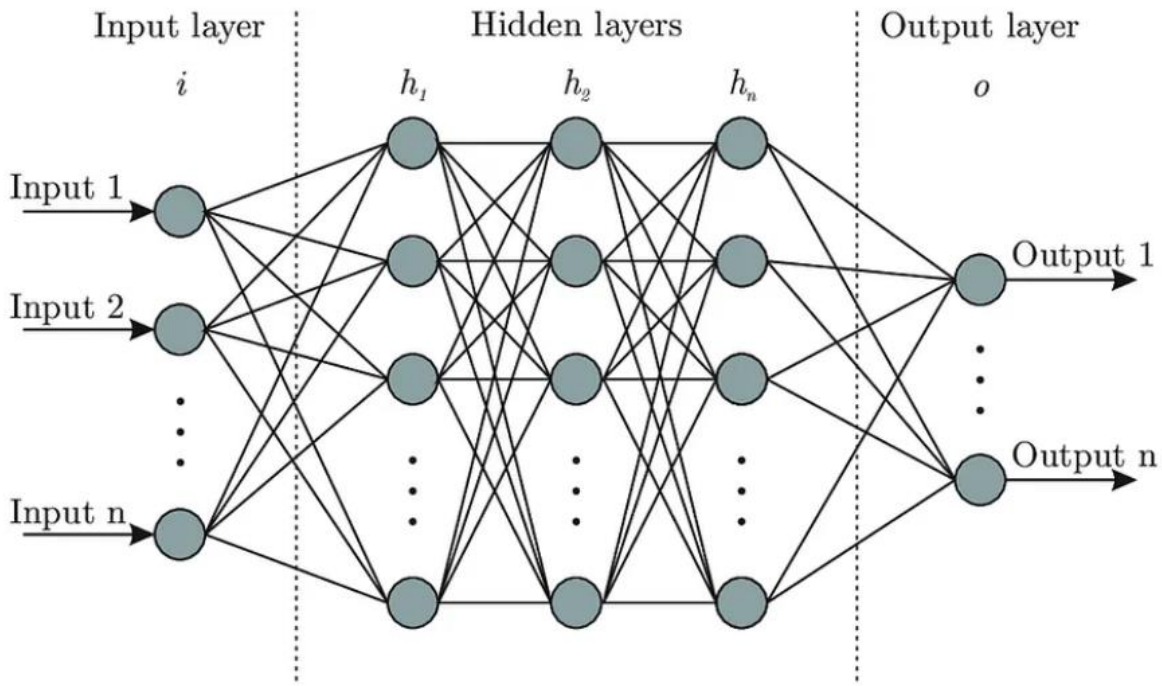

Figure 3 - Neural Network Architecture

B. NAIVE BAYES CLASSIFIER

Based on Bayesian theory, the Naive Bayes statistical method determines the outcome based on the probability with the highest likelihood. Using the known value as a basis, it calculates the probability of the unknown value [16]. It is possible to forecast an unknown probability by using reasoning and previous knowledge. Binary classes and conditional probabilities are the primary foundations of Naive Bayes.

$$\text{prob}\left(\text{class}_j|\text{feature}_k\right) = \frac{\text{prob}\left(\text{feature}_k|\text{class}_j\right) * \text{prob}\left(\text{class}_j\right)}{\text{prob}\left(\text{feature}_k\right)}, \quad (1)$$

$$\text{prob}\left(\text{feature}_k|\text{class}_j\right) = \prod_{j=1}^{m}\text{prob}\left(\text{feature}_k|\text{class}_j\right). \quad (2)$$

In equations (1) and (2), *n* indicates the maximum amount of features, prob(feature | class) indicates the probability of generating feature value feature provided in class, and prob(feature) and prob(class) indicate the probability of occurrence of feature value feature and the occurrence of class respectively. This classifier was utilized for binary classification with the aid of the Bayesian principle.

C. K-NEAREST NEIGHBOUR (KNN)

The KNN classifier is an example of a learning method wherein classification is done using the Minkowski distance function and the Manhattan or Euclidean measure of similarity. Although the Minkowski function works with categorical data, the Manhattan or Euclidean function mostly works with continuous variables. In the KNN classifier, the distance is calculated using the Euclidean function. Between two vectors (Xi and Xj), the Euclidean function (Dij) is computed using

$$\text{Dist}_{ij} = \sqrt{\sum_{l=1}^{m}\left(X_{il} - X_{jl}\right)^2}.$$

D. REGRESSION

To model the links between various variables and the probability of fraud, linear regression typically used to predict continuous outcomes—was modified [11]. By combining numerous input features at once, multiple regression expanded this and enabled a more in-depth examination of the ways in which different elements interacted to affect the likelihood of fraud.

In order to make sure the dataset was appropriate for regression analysis, data pretreatment included gathering patient records, billing data, and insurance claims. This was followed by feature engineering and cleaning. The multiple regression model offers a more comprehensive perspective by taking into account the combined influence of all relevant features.

The models were trained on this dataset. To make sure the models correctly reflected the underlying patterns, performance was assessed using measures like mean squared error and R-squared.

$$\text{Sig}_f(x) = \frac{1}{(1+e^{-x})},$$

$$x = w_0z_0 + w_1z_1 + w_2z_2 + \cdots + w_nz_n.$$

V. DATA AND DESCRIPTIVE STATISTICS

The dataset used for this work is from Kaggle datasets of health insurance medical records. The data description and various data pre-processing performed on the data are presented here. There are a total of 2,84,807 rows and 31 columns. The various data pre-processing performed on this dataset is described in this section.

A. MISSING VALUE ANALYSIS

The data used consists of claim records of individuals. Figure 4 shows the heatmap of the missing values in the dataset. The features which had a very high percentage of missing values were not considered for this study.

The heatmap of the dataset after removing features with a very high percentage of missing values is shown in Figure 4. Remaining features with few missing values in the dataset were handled separately. Other missing values in the dataset were handled using statistical methods.

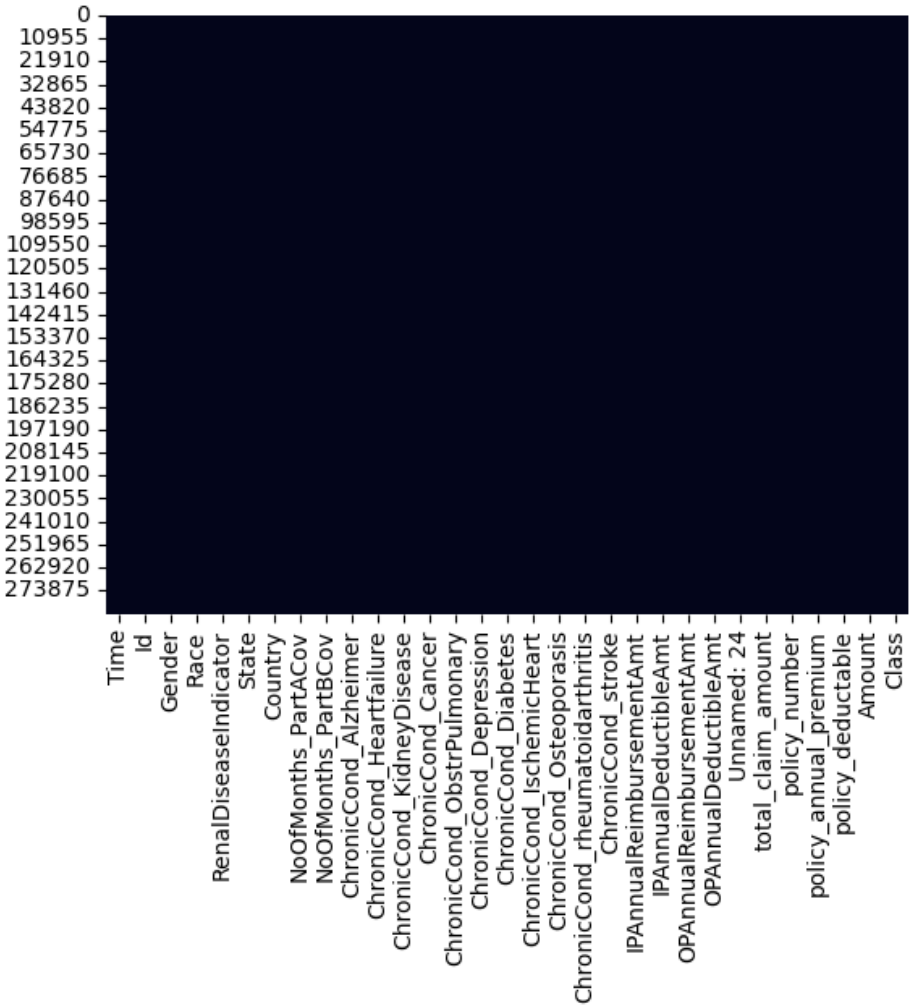

Figure 4 – Heatmap of the resulting dataset

B. EXPLORATORY DATA ANALYSIS

Figure 5 shows the histogram of the dataset for the numerical fields in the dataset.

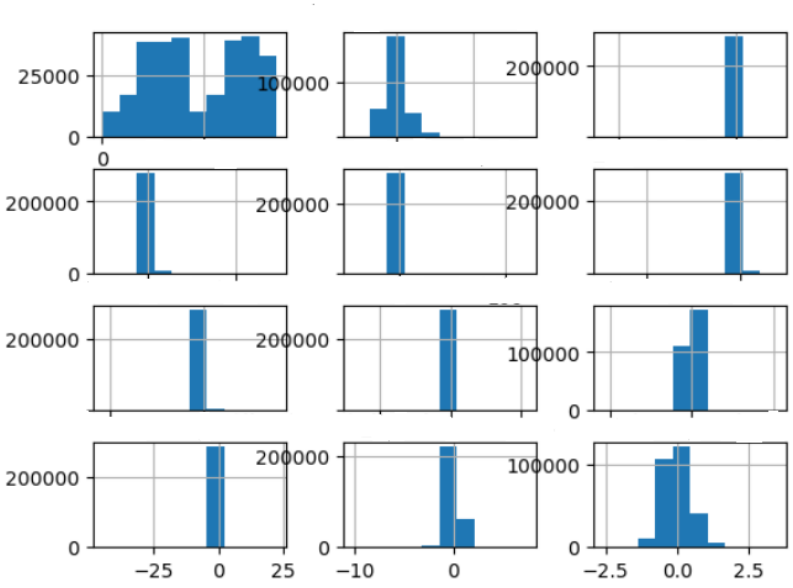

Figure 5 – Histogram of various fields of the dataset

C. CORRELATION GRAPH

Figure 6 shows the correlation matrix between the variables in the dataset. Some features which were highly correlated with each other were removed, and only one of them was considered for the work [13]. For example, ‘A’ is highly correlated with ‘B’, ‘C’, ‘D’. In this situation, ‘A’ is considered, and others were removed.

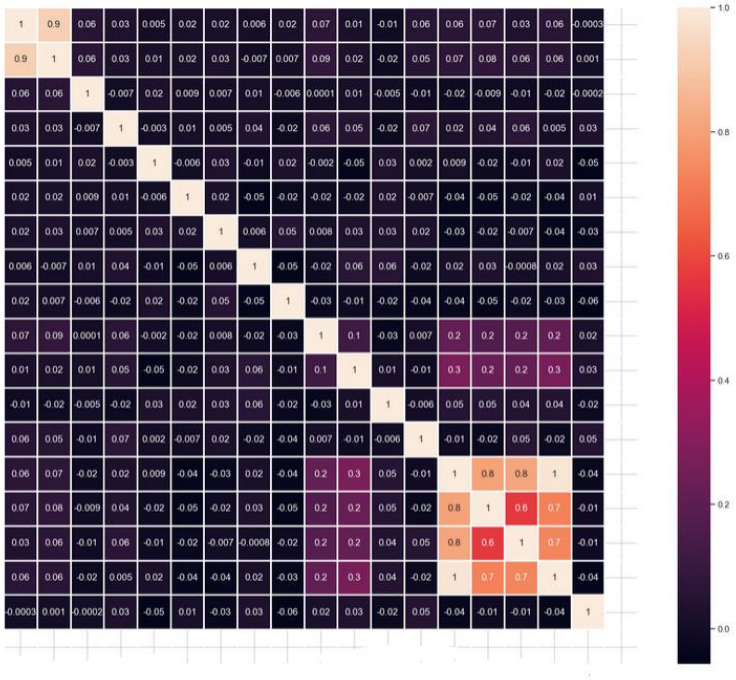

Figure 6 – Correlation graph of all the features in the dataset

All the highly correlated features in the dataset will be removed. Figure 7 shows the correlation matrix after removing the highly correlated features in the dataset. We can observe that some features which are least correlated are not removed.

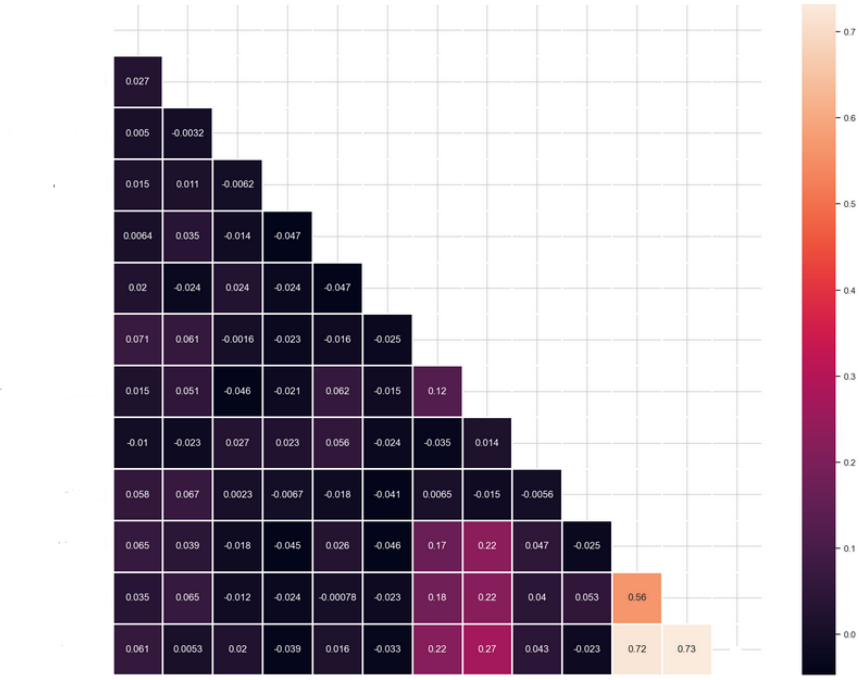

Figure 7 – Correlation graph after removing highly correlated features

D. OUTLIERS

Figure 8 displays the outliers of the dataset. Machine learning model's performance is affected by the outliers. To detect outliers, a variety of models are used, including Z-scores, statistical methods, and visualization techniques.

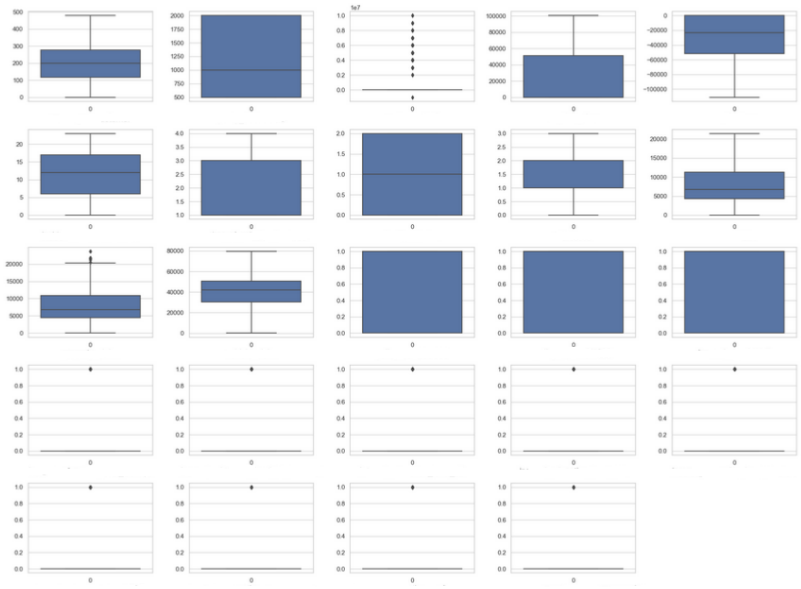

Figure 8 – Outliers of the different fields in the dataset

VI. RESULTS

This section contains the results of various studies that were carried out. The dataset used was divided inti train and test in the ratio 80:20. The dataset is highly imbalanced in nature, with 25% of the claim as fraudulent, which is 71,200 claims out of 2,84,807. Figure 9 shows the histogram of fraud and not-fraud claims in the dataset.

Figure 9 – Histogram of fraud and not-fraud claims in the dataset

We have used different classification models and implemented them on the dataset. Thus, for each of the model, there are different outputs. Table 3 contains the performance metrics for this study [14]. Different metrics are used for comparing the models – accuracy, precision, recall, and F1-score. For each of the metrics, the highest scores are highlighted. However, it is essential to check the performance of the model, considering all the metrics.

- **Accuracy:** It indicates the degree to which the measured value resembles known value.
- **Precision**: This indicates the model's level of accuracy with respect to positively anticipated data.
- **Recall:** After classifying something as positive, it determines how many real positives the model caught (true positives).
- **F1-score:** It provides a recall and accurate balancing.

$$\textbf{Accuracy} = (TP + FN)/ (TP + TF + FP + FN)$$

$$\textbf{Precision} = TP / (TP + FP)$$

$$\textbf{Recall} = TP / (TP + FN)$$

$$\textbf{F1} = 2 \times (\text{Recall} \times \text{Precision}) / (\text{Recall} + \text{Precision})$$

Here,

TP: True Positive

FP: False Positive

TN: True Negative

FN: False Negative

Models	Accuracy	Precision	Recall	F1-score	Support
SVM	**0.87**	0.74	**1.00**	**0.85**	**186**
KNN	0.76	0.74	**1.00**	**0.85**	**186**
Ada Boost	0.81	0.86	0.72	0.78	**186**
Decision Tree	1.00	0.79	0.50	0.61	**186**
Naïve Bayes	0.84	**1.00**	0.67	0.80	003
Voting Classifier	0.82	0.74	**1.00**	**0.85**	**186**

Table 3 – Performance metrics of different Machine Learning models

Thus, it is essential to find other models which had an improved performance while having better scores with respect to other metrics. For this, deep learning models were implemented to identify a more efficient model.

CNN is a type of deep learning technique and has layers. The convolutional layers apply operations using a set of filters to the features from the input. Table 4 represents the architecture of the convolutional neural network model used in this study.

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 30, 32)	96
batch_normalization (BatchNormalization)	(None, 30, 32)	128
max_pooling1d (MaxPooling1D)	(None, 15, 32)	0
dropout (Dropout)	(None, 15, 32)	0
conv1d_1 (Conv1D)	(None, 15, 64)	4,160
batch_normalization_1 (BatchNormalization)	(None, 15, 64)	256
max_pooling1d_1 (MaxPooling1D)	(None, 7, 64)	0
dropout_1 (Dropout)	(None, 7, 64)	0
flatten (Flatten)	(None, 448)	0
dense (Dense)	(None, 64)	28,736
dropout_2 (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 1)	65

Total params: 33,441 (130.63 KB)

Trainable params: 33,249 (129.88 KB)

Non-trainable params: 192 (768.00 B)

Table 4 – Architecture of the Convolutional Neural Network Model

RNN is a type of neural network used to recognize patterns in data and it also has layers. These layers process the sequence data. Each unit in an RNN layer maintains the information from previous time steps. In this study, we used Long Short-term Memory (LSTM) to capture better long-term dependencies in the sequence data. Table 5 represents the architecture of the recurrent neural network model used in this study.

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 5, 1)	12
lstm_1 (LSTM)	(None, 5, 1)	12
lstm_2 (LSTM)	(None, 5, 1)	12
lstm_3 (LSTM)	(None, 1)	12

Total params: 48 (192.00 B)

Trainable params: 48 (192.00 B)

Non-trainable params: 0 (0.00 B)

Table 5 – Architecture of the Recurrent Neural Network Model

The obtained results were plotted to visualize the comparison in terms of performance metrics [12]. First, training accuracy vs. validation accuracy is represented in Figure 10. Second, training loss vs. validation loss is represented in Figure 11.

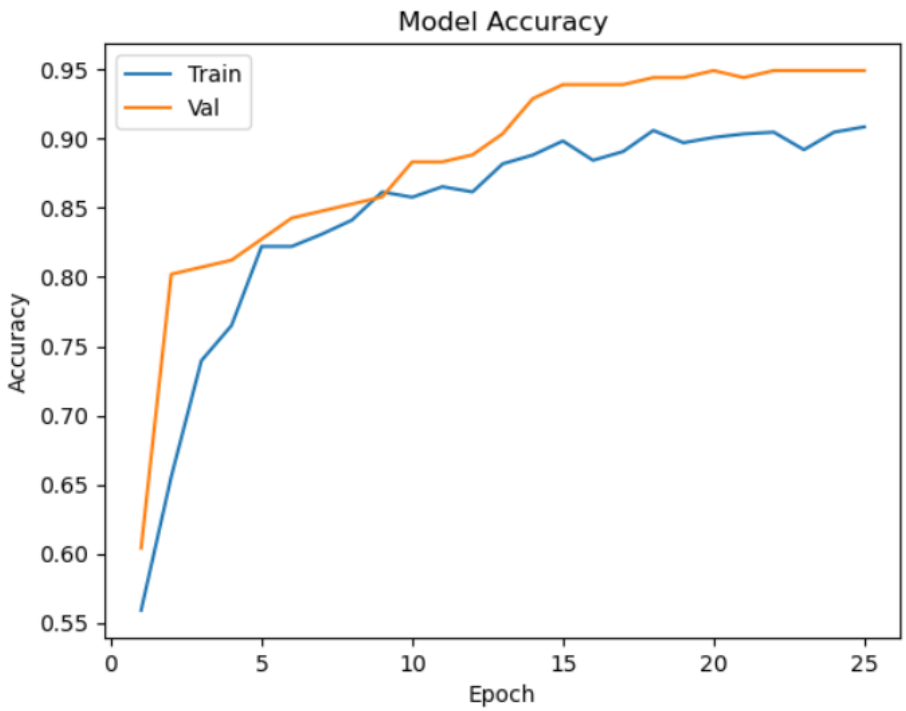

Figure 10 – Representation of training accuracy vs. validation accuracy

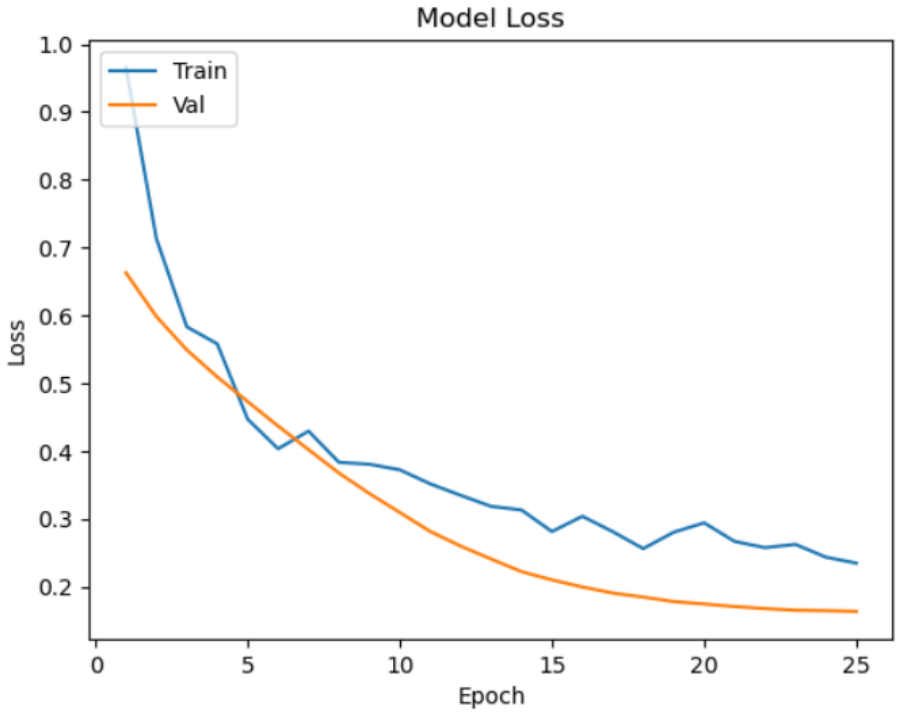

Figure 11 – Representation of training loss vs. validation loss

VII. CONCLUSION

Our goal in this research is to develop a strong fraud detection model for the healthcare system by doing a thorough analysis of several machine learning and deep learning models. It was observed that the neural networks, when trained on the dataset, performed better than other classification models. The neural network model has the greatest F1-score of 0.95, which suggests a trustworthy detection capability and a balanced trade-off between precision and recall.

The success of neural networks can be attributed to their ability to spot complex relationships and patterns in the data that traditional machine learning models usually miss. This demonstrates the critical role deep learning approaches play in handling the complicated, high-dimensional datasets that are commonly used in healthcare fraud investigation.

In summary, neural networks can be an effective weapon in the battle against healthcare fraud if they are combined with effective techniques for correcting data inconsistencies. The promising results suggest that applying similar approaches to different business domains could help identify the best machine learning or deep learning models for fraud detection.

To determine which deep learning or machine learning models work best, a similar study can be conducted in the future with additional business domains.

REFERENCES

[1] Mehbodniya, Abolfazl, et al. "[Retracted] Financial Fraud Detection in Healthcare Using Machine Learning and Deep Learning Techniques." *Security and Communication Networks* 2021.1 (2021): 9293877.

[2] Gupta, Rohan Yashraj, Satya Sai Mudigonda, and Pallav Kumar Baruah. "A comparative study of using various machine learning and deep learning-based fraud detection models for universal health coverage schemes." *International Journal of Engineering Trends and Technology* 69.3 (2021): 96-102.

[3] Thornton, Dallas, et al. "Categorizing and describing the types of fraud in healthcare." *Procedia Computer Science* 64 (2015): 713-720.

[4] Kelley, Robert. "Where can $700 billion in waste be cut annually from the US healthcare system." *Ann Arbor, MI: Thomson Reuters* 24 (2009): 1-30.

[5] Koh, Hian Chye, and Gerald Tan. "Data mining applications in healthcare." *Journal of healthcare information management* 19.2 (2011): 65.

[6] Huang, Hongcheng, and Ziyu Dong. "Research on architecture and query performance based on distributed graph database Neo4j." *2013 3rd International Conference on Consumer Electronics, Communications and Networks*. IEEE, 2013.

[7] Saldamli, Gokay, et al. "Health care insurance fraud detection using blockchain." *2020 seventh international conference on software defined systems (SDS)*. IEEE, 2020.

[8] Duman, Ebru Aydoğan, and Şeref Sağıroğlu. "Heath care fraud detection methods and new approaches." *2017 International Conference on Computer Science and Engineering (UBMK)*. IEEE, 2017.

[9] Lo'Ai, A. Tawalbeh, and Suhaila Habeeb. "An integrated cloud-based healthcare system." *2018 fifth international conference on internet of things: systems, management and security*. IEEE, 2018.

[10] Gupta, Rohan Yashraj, Satya Sai Mudigonda, and Pallav Kumar Baruah. "A comparative study of using various machine learning and deep learning-based fraud detection models for universal health coverage schemes." *International Journal of Engineering Trends and Technology* 69.3 (2021): 96-102.

[11] Bauder, Richard A., and Taghi M. Khoshgoftaar. "Medicare fraud detection using machine learning methods." *2017 16th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, 2017.

[12] Gupta, Rohan Yashraj, et al. "Implementation of a predictive model for fraud detection in motor insurance using gradient boosting method and validation with actuarial models." *2019 IEEE international conference on clean energy and energy efficient electronics circuit for sustainable development (INCCES)*. IEEE, 2019.

[13] Gupta, Rohan Yashraj, et al. "Implementation of correlation and regression models for health insurance fraud in Covid-19 environment using actuarial and data science techniques." *arXiv preprint arXiv:2102.04210* (2021).

[14] Lu, Fletcher, and J. Efrim Boritz. "Detecting fraud in health insurance data: Learning to model incomplete Benford's law distributions." *European Conference on Machine Learning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.

[15] Matloob, Irum, and Shoab Khan. "A framework for fraud detection in government supported national healthcare programs." *2019 11th international conference on electronics, computers and artificial intelligence (ECAI)*. IEEE, 2019.

[16] Agarwal, Shashank. "An Intelligent Machine Learning Approach for Fraud Detection in Medical Claim Insurance: A Comprehensive Study." *Scholars Journal of Engineering and Technology* 11.9 (2023): 191-200.