Grenze International Journal of Engineering and Technology, Jul Issue

Fraud Detection in Health Insurance Claims using Machine Learning and Deep Learning Techniques

Akshatha P[1], Manasa G L[2], Nayanashree H C[3] and Dr. A S Poornima[4]
[1-4]Department of Computer Science Engineering, Siddaganga Institute of Technology
Tumkur, India

***Abstract*—A claim is what a doctor or hospital submits to the insurance company so they can get paid. A health insurance claim is a request that a health insurance policyholder submits to the Insurance Company in order to get the services that are covered in their health insurance policy.To tackle the problem of fraud in the medical insurance claims, health insurance companies make use of traditional rule-base models, but these models do not suffice anymore due to several factors such as the large volume of claims to be processed which makes the medical billing process prone to error, slow and sometimes inefficient. Hence there is a need to develop a novel fraud detection model for insurance claims processing. In this paper we are using the power of Machine Learning(ML) and Deep Learning(DL) algorithms to efficiently identify the fraudulent health insurance claims from the healthcare transaction dataset. Here we have considered a data set consisting of both legitimate and fraudulent claims. Initially we have implemented ML algorithms like SVM(Support Vector Machine), Logistic regression and Random forest for detection of anomalies and classification of health insurance claims into legitimate and fraudulent claims. Also DL algorithm MLP (Multilayer perceptron) is implemented. We have measured various metrics like Accuracy, Recall, Precision etc., for evaluating the performance of these ML and DL algorithms. All the models are good, but comparatively MLP has better performance with highest accuracy of 88.6 %. Study is also performed by varying the number of fraudulent records for the initial data set. This model helps in easy detection of the occurrence of fraudulent activities also is capable of fair fraud identification when applied to datasets with different fraud occurrence rates. The model can be used for decreased fraudulent activities in health insurance claims, it can become an efficient and effective application for Insurance clients and companies.**

***Index Terms*— Insurance Frauds, Fraud Detection, Machine Learning, Support Vector Machine, Logistic regression, Random forest, Multilayer perceptron.**

I. INTRODUCTION

The health insurance industry, a foundation in modern-day society, serves the purpose of providing individuals with affordable healthcare [4]. Healthcare fraud happens when someone like healthcare provider or an individual intentionally provides false information or withholds information in order to get a financial benefit. Common types of healthcare fraud are identity theft, charging for services that weren't provided and falsifying documents to make claims.Healthcare fraud costs about hundreds of millions of dollars loss each year to private health insurance industry.Insurance fraud has led to significant negative impacts to the insurance sector for several decades. These losses drive up premiums and make the system less efficient.

Grenze ID: 01.GIJET.7.2.11

In health insurance the risk of unnecessary or non-existent medical services due to misrepresentations by patients or providers is a costly one. By improving access to resources for urgent and chronic care rather than wasting time and funds seeking out bad actors, clinicians and medical consultants can fundamentally improve the quality of care they provide using data science, machine learning, and AI [1].

Machine learning is the study of computer algorithms that improve automatically through experience. It is seen as a sub substitute part of artificial intelligence. Machine learning algorithm builds a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. The term machine learning was invented by Arthur Samuel in 1959, an American IBMer. He is also the pioneer of computer gaming and artificial intelligence[2].

To tackle the problem of fraud in the medical insurance claims, health insurance companies make use of traditional rule-base models, but these models are not adequate due to several factors such as the large volume of claims to be processed which makes the medical billing process prone to error, slow and sometimes inefficient. There is a growing concern in the insurance field about the increasing incidence of abuse and fraud in health insurance. Hence there is a need to develop a novel fraud detection model for insurance claims processing.

In the proposed system, we are using the power of machine learning and deep learning algorithms to efficiently identify the fraudulent health insurance claims from the healthcare transaction dataset. The initial step is dataset collection and preparation. We are taking the 3 datasets with increasing fraud rates. The data preprocessing stage involves the removal of unwanted customers and missing records. We are using the SVM(Support Vector Machine), Logistic regression, random forest and MLP classifier for detection of anomalies and classification of health insurance claims into legitimate and fraudulent claims. We are going to compare the accuracy of each algorithm among three versions and also compare the accuracy of each algorithm with another.By making use of health insurance fraud claim data, machine learning algorithms and deep learning techniques, we can detect the occurrence of fraudulent activities in effective and efficient manner.

All the models proposed in this paper are good, the performance metrics are recorded and it shows that even with the increased fraudulent records in the data set, MLP has better and steady performance with highest accuracy of 88.6 % compared to other algorithms. The proposed model helps in easy detection of the occurrence of fraudulent activities also is capable of fair fraud identification when applied to datasets with different fraud occurrence rates. The model can be used for decreased fraudulent activities in health insurance claims, it can become an efficient and effective application for Insurance clients and companies.

II. Literature Survey

To Understand which field and method is best to detect healthcare frauds. And also which all machine learning algorithms are feasible, numerous literatures and theories were proposed already and are in public usage. “A Secure AI-Driven Architecture for Automated Insurance Systems: Fraud Detection and Risk Measurement” by NajmeddineDhieb et al. [3] describes the development of a secure and automated insurance system framework which reduces human interaction, secures the insurance activities, alerts and informs about risky customers, detects fraudulent claims, and reduces monetary loss for the insurance sector. It presents a novel Smart Insurance System based on Blockchain and Artificial intelligence to codify business rules, automate claims processing, estimate clients risk levels, and detect fraudulent claims. In this they proposed a unified architecture which uses permissioned blockchain as a secure network for insurance companies to store and share different information about customers and claims. They have proposed two machine learning algorithms: Extreme gradient boosting (XGBoost) and Very fast decision tree to build the fraud detection and risk measurement modules.

Another literature in this domain is “Framework for Analysis and Detection of Fraud in Health Insurance” by Nirmal Rayan [4]explores the use of statistical methods to create a rule based engine that works with self-learning Decision trees.It introduces a hybrid framework that combines domain expertise, supervised learning and unsupervised learning techniques to identify fraudulent claims from a given set of outstanding claims.

The similar research in the field of healthcare fraud detection is “Decision Support System (DSS) for Fraud Detection in Health Insurance Claims Using Genetic Support Vector Machines (GSVMs)” by Robert A. Sowahet al.[5].The aim of this paper is to develop a novel fraud detection model for insurance claims process which is based on genetic algorithms and support vector machine. GSVM is used for anomaly detection and classification of health insurance claims into legitimate and fraud claims. This paper also gives the theoretical and mathematical foundations of genetic algorithms (GA), support vector machines (SVMs), and the hybrid genetic support vector machines (GSVMs) in combating the global phenomenon.

The other alternate method proposed for our approach is “A Comparison of Machine Learning Methods Applicable to Healthcare Claims Fraud Detection” by NnaemekaObodoekwe and Dustin Terence van der Haar

[6] implemented a prototype that comprised different methods and a comparison of each of the methods was carried out to determine which method is most suited for the Medicare dataset. The study explored the application of different machine learning methods like Bayesian classifier, Logistic regression, Random forest classifier Artificial Neural Network to detect fraudulent claims in the medical billing process. The result generated from the application of these machine learning methods were collected.
In this paper we aim at not only machine learning algorithms but also deep learning. Also in our study we are creating data sets consisting of varied number of fraudulent records to study the performance of the various algorithms.

III. PROPOSED SYSTEM

In this section we are going to explain about data set preparation and algorithms implemented.

A. Dataset preparation:

We are taking the 3 datasets with increasing fraud rates. Initial dataset i.e., dataset 1 is taken from kaggle site. Dataset 2 is prepared by adding 100 more fraud cases to dataset 1. Similarly dataset 3 is prepared by adding 100 more fraud cases to dataset 2. Each of the dataset has the following files:
• Inpatient Data: This data provides insights about the claims for those patients who are admitted to the hospitals. It provides details like their admission and discharge date, diagnosis code.
• Outpatient Data: This data provides details about the claims for those patients who visit hospitals and are not admitted. • Beneficiary Data: This data contains details of patients like health condition, address, etc.
The dataset we have considered is real world data and it is in structured format. Normally, data in the real world is dirty, contains missing values and can be incorrect. Therefore, a lot of cleaning up activity has to be done on the data to get it to the form that will be suitable for use.The data pre-processing carried out by removing unwanted fields and adding required values. Since the variable factors considered for dataset are highly confidential, we do not reveal those variables and apply algorithms for the dataset.
We applied 75:25 split to the dataset. 75% of the data is used for training the machine learning model and 25% used for testing the outcome.

B. Algorithms implemented:

After completing the pre-processing and transformation of the data, we apply following algorithms to derive insights from the data.
1. Logistic regression: Logistic regression is a statistical model and it uses logistic function to model a binary dependent variable. In regression analysis, it is estimating the parameters of a logistic model(a form of binary regression) [7].
2. Support Vector Machine (SVM) : It is a supervised machine learning algorithm which can be used for both classification or regression challenges. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine [8].
3. Random Forest: It is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Random Forest works in two-phase. First step is to create the random forest by combining N decision tree, and second step is to make predictions for each tree created in the first phase [9].
4. Multilayer perceptron (MLP) : It is a deep, artificial neural network. It is composed of more than one perceptron. They are composed of three layers namely input layer, hidden layer, output layer. Input layer receive the signal and output layer makes a decision or prediction about the input. In between those two, an arbitrary number of hidden layers that are the true computational engine of the MLP will be there. MLPs with one hidden layer are capable of approximating any continuous function[10].
A detailed flow of the process is described with a flow diagram in Figure 1.

IV. RESULT ANALYSIS

A Classification report measures the quality of predictions from a classification algorithm. It shows main classification metrics like precision, recall and f1-score on a per-class basis. The metrics are calculated by using true positive, false positive, true negative and false negative count..In the proposed system, we have considered accuracy as a measuring metric as in thiscase it is important to predict the fraud cases as fraud and non fraud cases as genuine claims. Following are the metrics considered for evaluating the implemented algorithms.

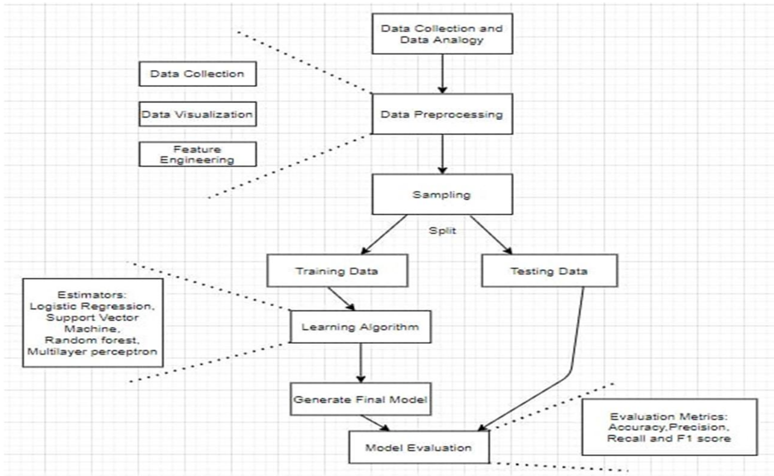

Figure 1. Flow diagram

Accuracy –It is defined as the ratio of correctly predicted observation to total number of observations. Accuracy can tell us immediately whether a model is being trained correctly and how it may perform generally [11].
Accuracy = TP+TN/TP+FP+FN+TN.

Precision - Precision is defined as the ratio of positive observations which are predicted correctly to the total predicted positive observations. High precision relates to the low false positive rate [11].Precision = TP/TP+FP

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class – yes [11]. Recall = TP/TP+FN

F1 score – It is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account [11]. F1 Score = 2*(Recall * Precision) / (Recall + Precision).

Observations and results are recorded for *Dataset1, Dataset2 and Dataset3* separately. Following subsections describe the results in detail.

A. Results and observations for Dataset1

After pre-processing number of instances obtained: 5410
Number of instances considered for testing: 1353
Number of Fraud Transactions :135
Number of Valid Transactions : 1218
Table I shows classification report for Dtaset1. Accuracy score and respective confusion matrix are tabulated in Table II.

TABLE I. CLASSIFICATION REPORT FOR DATASET 1

Algorithm		Precision	Recall	F1-Score	Support
Logistic Regression	0	0.99	0.83	0.90	1233
	1	0.34	0.92	0.50	120
	Average/Total	0.93	0.84	0.87	1353
Support Vector Machine	0	0.99	0.85	0.91	1233
	1	0.36	0.88	0.51	120
	Average/Total	0.93	0.85	0.88	1353
Random Forest Classifier	0	0.99	0.77	0.87	1233
	1	0.28	0.93	0.43	120
	Average/Total	0.93	0.78	0.83	1353
Multi Layer Perceptron Classifier	0	0.97	0.94	0.95	1233
	1	0.52	0.68	0.59	120
	Average/Total	0.93	0.92	0.92	1353

B. Results and observations for Dataset2

Number of instances obtained: 5512
Number of instances considered for testing: 1378
Number of Fraud Transactions : 165
Number of Valid Transactions : 1213

TABLE II. CONFUSION MATRIX AND ACCURACY SCORE FOR DATASET 1

Algorithm	Confusion Matrix	Accuracy Score
Logistic regression	$\begin{bmatrix} 1023 & 210 \\ 10 & 110 \end{bmatrix}$	83.74%
Support Vector Machine	$\begin{bmatrix} 1046 & 187 \\ 15 & 105 \end{bmatrix}$	85.07%
Random Forest Classifier	$\begin{bmatrix} 950 & 283 \\ 9 & 111 \end{bmatrix}$	78.42%
Multi Layer Perceptron Classifier (MLP)	$\begin{bmatrix} 1158 & 75 \\ 38 & 82 \end{bmatrix}$	91.65%

Table III shows classification report for Dtaset1. Accuracy score and respective confusion matrix are tabulated in Table IV.

TABLE III. CLASSIFICATION REPORT FOR DATASET 2

Algorithm		Precision	Recall	F1-Score	Support
Logistic Regression	0	0.96	0.83	0.89	1226
	1	0.34	0.92	0.50	152
	Average/Total	0.89	0.81	0.84	1378
Support Vector Machine	0	0.96	0.85	0.90	1226
	1	0.37	0.74	0.49	152
	Average/Total	0.90	0.83	0.86	1378
Random Forest Classifier	0	0.98	0.40	0.57	1226
	1	0.16	0.93	0.28	152
	Average/Total	0.89	0.46	0.54	1378
Multi Layer Perceptron Classifier	0	0.95	0.94	0.94	1226
	1	0.55	0.60	0.57	152
	Average/Total	0.91	0.90	0.90	1378

TABLE IV. CONFUSION MATRIX AND ACCURACY SCORE FOR DATASET2

Algorithm	Confusion Matrix	Accuracy Score
Logistic regression	$\begin{bmatrix} 1014 & 212 \\ 43 & 109 \end{bmatrix}$	81.49%
Support Vector Machine	$\begin{bmatrix} 1037 & 189 \\ 40 & 112 \end{bmatrix}$	83.38%
Random Forest Classifier	$\begin{bmatrix} 496 & 730 \\ 11 & 141 \end{bmatrix}$	46.23%
Multi Layer Perceptron Classifier (MLP)	$\begin{bmatrix} 1152 & 74 \\ 61 & 91 \end{bmatrix}$	90.20%

C. Results and observations for Dataset3:

Number of instances obtained: 5613
Number of instances considered for testing: 1404
Number of Fraud Transactions : 196
Number of Valid Transactions : 1208
Table V shows classification report for Dtaset3. Accuracy score and respective confusion matrix are tabulated in Table VI.

TABLE V. CLASSIFICATION REPORT FOR DATASET3

Algorithm		Precision	Recall	F1-Score	Support
Logistic Regression	0	0.93	0.82	0.87	1229
	1	0.32	0.59	0.41	175
	Average/Total	0.86	0.79	0.82	1404
Support Vector Machine	0	0.96	0.84	0.90	1229
	1	0.40	0.75	0.53	175
	Average/Total	0.89	0.83	0.85	1404
Random Forest Classifier	0	0.98	0.41	0.58	1229
	1	0.18	0.93	0.31	175
	Average/Total	0.88	0.47	0.54	1404
Multi Layer Perceptron Classifier	0	0.93	0.94	0.93	1229
	1	0.54	0.54	0.54	175
	Average/Total	0.89	0.89	0.98	1404

TABLE VI. CONFUSION MATRIX AND ACCURACY SCORE FOR DATASET3

Algorithm	Confusion Matrix	Accuracy Score
Logistic regression	$\begin{bmatrix} 1009 & 220 \\ 72 & 103 \end{bmatrix}$	79.20%
Support Vector Machine	$\begin{bmatrix} 1036 & 193 \\ 44 & 131 \end{bmatrix}$	83.12%
Random Forest Classifier	$\begin{bmatrix} 501 & 728 \\ 12 & 163 \end{bmatrix}$	47.29%
Multi Layer Perceptron Classifier (MLP)	$\begin{bmatrix} 1150 & 79 \\ 81 & 94 \end{bmatrix}$	88.60%

Comparison graph: Figure 2 shows the comparison of performance of algorithms with respect to each of different datasets. After analysis it is observed that MLP Classifier is consistent over all the datasets considered.

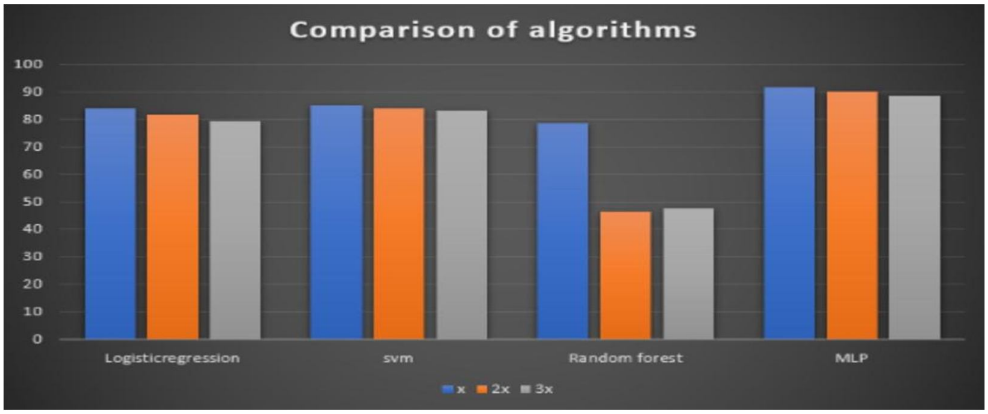

Figure 2. Comparison of algorithms

We can also improve on this accuracy by increasing the sample size, however at the cost of computational expense we can also complex anomaly detection models to get better accuracy in determining more fraudulent cases.

V. CONCLUSION

The task of building a system that identifies the possible fraudulent healthcare claims has become very important as the demand for healthcareincreases. To identify the fraud in the medical insurance claims, health insurance companies make use of traditional rule-base models. But they are not efficient. ML and DL algorithms have improved the effectiveness of insurance claim process.

This paper gives the insight about three ML and one DLalgorithms that is, Logistic regression, Support vector machines, Random forest classifier and Multilayer perceptron classifier forfraud detection. These four algorithms are applied to three different datasets with increasing fraud rates and their accuracies have been recorded. After the analysis of result it is observed that, all the models are good, but comparatively MLP has better performance with highest accuracy of 88.6 %. It is also observed that the performance of SVM, Logistic regression and MLP is consistent over all 3 datasets. Random forest classifier gives great accuracy when applied to dataset with minimum fraud rates but fails miserably when fraud rates are increased.

This model helps in easy detection of the occurrence of fraudulent activities also is capable of fair fraud identification when applied to datasets with different fraud occurrence rates. The model can be used for decreased fraudulent activities in health insurance claims, it can become an efficient and effective application for Insurance clients and companies.

For future study we can use the live or actual data from insurance company or hospital for the real time study.

REFERENCES

[1] “Fraud-detection-in-healthcare”,[Online]. Available: “https://www.dataiku.com/stories/machine-learning-based-fraud-detection-in-healthcare/”.

[2] “Machine learning – Wikipedia”,[Online]. Available: “https://en.wikipedia.org/wiki/Machine_learning”.
[3] NajmeddineDhieb, Hakim Ghazzai, HichemBesbes and Yehia Massoud, “A Secure AI-Driven Architecture for Automated Insurance Systems: Fraud Detection and Risk Measurement”, IEEE Access.
[4] Nirmal Rayan, “Framework for Analysis and Detection of Fraud in Health Insurance”, Proceedings of CCIS2019.
[5] Robert A. Sowah, Marcellinus Kuuboore, Abdul Ofoli, Samuel Kwofie, Louis Asiedu, Koudjo M. Koumadi and Kwaku O. Apeadu, “Decision Support System (DSS) for Fraud Detection in Health Insurance Claims Using Genetic Support Vector Machines (GSVMs)”, Hindawi Journal of Engineering.
[6] NnaemekaObodoekwe and Dustin Terence van der Haar, “A Comparison of Machine Learning Methods Applicable to Healthcare Claims Fraud Detection”, Springer Nature Switzerland AG 2019.
[7] “Logistic regression”, Available: “https://en.m.wikipedia.org/wiki/Logistic_regression”.
[8] “Support-Vector-Machine”,Available:”https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm”.
[9] "Machine Learning Random Forest Algorithm - Javatpoint" https://www.javatpoint.com/machine-learning-random-forest-algorithm.
[10] "A Beginner's Guide to Multilayer Perceptrons (MLP) | Pathmind" https://wiki.pathmind.com/multilayer-perceptron.
[11] https://medium.com/nerd-for-tech/confusion-matrix-demystify-the-confusion-5d7f4801ad4f.