Grenze International Journal of Engineering and Technology, June Issue

A Comparative Analysis for Identifying the Fraudulent Healthcare Claims

Dr. P. Naga Jyothi[1], Dr. M Venkata Ramana[2] and Dr. S. Suresh[3]
[1-2]Dept. of CSE, GITAM School of Technology, GITAM University, Visakhapatnam, INDIA
Email: pbtjyothiraj.33@gmail.com, vmancha@gitam.edu
[3]Chief Librarian, ANITS, Tagarapuvalasa, Bheemili,Visakhapatnam, INDIA
Email: sureshs.nice@gmail.com

***Abstract*—Medical expenses have sharply increased in recent times, primarily due to the rise in the elderly population. To effectively manage resources for all groups of people, health insurance policies have been developed. These programs can help educate citizens about cost control, quality fees, expenses, budgeting, and future payments, enabling them to utilize healthcare services at affordable costs. Supervised machine learning algorithms are playing a dominant role in identifying fraudulent claims submitted by hospitals, patients, or doctors. This is significantly reducing the financial burden on the public, the government, and private health insurance providers. To prevent mismanagement of costs in the healthcare sector, there is a growing demand to identify falsified (fraudulent) claims. A considerable amount of resources is being spent on developing fraud control systems to deal with fraudulent practitioners and planned illegal patterns. In this study, a comparative analysis using machine learning algorithms like Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), and Artificial Neural Networks (ANN) was conducted to analyze behaviors and identify fraudulent and non-fraudulent claims in the Center for Medicaid and Medicare Services (CMS) data. The results show that SVM had superior accuracy compared to other algorithms. The analysis also includes different metrics to evaluate relative performance. Overall, the development and implementation of effective fraud control systems and the use of advanced machine learning algorithms can help improve the management of healthcare costs and ensure that resources are utilized efficiently and effectively.**

***Index Terms*— medical expenditures, supervised learning, machine learning, fraudulent behaviors**.

I. INTRODUCTION

Fraud is an eclectic spread term in most applications. Fraud is defined as intentional deception to misrepresent the result for an unauthorized benefit. According to the global economic crime survey by PwC in 2018 that 49 percent of organizations suffer from fraud. Every year this count increases to billions–of dollars in business. The possibilities of fraud are co-evolving with technology. There is a fable that 'fraud in insurance is victimless' day by day, it is increasing and has become a massive problem for all the insurance companies. Insurers invest money in claims analysis and spend colossal effort and workforce in detecting and preventing fraud. Fraud analytics is a probe into digital technology across industries. Fraud analytics helps in recognizing solutions at an early stage before it results in a loss. It severely spoils the assurance ecosystem in various aspects like damage to

Grenze ID: 01.GIJET.9.2.812_1

reputation, general costs, governing compliance, unfairness with rightfully deserving, endorsing, and loss of faith. Always there is a challenge to detect and deal with fraud. They must rely on experts, adjusters, and special investigators for justification to agents. The fraud management platform advances the bottom line by mitigating fraud, abuse, and waste. The system helps process excessive data and combines structured, unstructured, and other sources [1][2][3].A claim is a formal request for money after the primary loss, and it helps the people in the society from financial ruin. There are various claims like Business Insurance Claims, Fire & Smoke Damage Claims, Hail Damage Claims, Health Insurance claims, Homeowner Insurance Claims, etc.,

A. Fraud in Insurance Claims

This paperwork focuses on health insurance claims as healthcare data grew from 2012 to 2020 by 50 percent, i.e., predicted as 2,314 Exabytes (nearly 25,000 petabytes) drastically doubles every 73 days. In the National Health Expenditure data (NHE) report, the average rate of national health expenses is likely to grow from 5.5 percent to $6.0 trillion by 2027. The Gross Domestic Product (GDP) is expected to rise from 0.8 percent to 17.9 percent of the US government. The critical factors are economic and demographic to the health sector during 2018-2027. Healthcare has a diversified nature of data which are as follows: Administrative enrollment data, Billing records data(charges and payments), Medical records data, Hospitals management data, Physicians related data, Health plans information(private/public), Procedures data(diagnosis), Patient-related data(in/out), Treatment-related data(clinical, symptoms, disease), and Demographic details(patient/hospital)[4].Fraud prevention and detection have to work simultaneously to stop the occurrence of fraud and handle it in the best possible way. In practice, detecting fraud starts with preventing false alarms, like identifying the tradeoff between the costs of analyzing a fraudulent case and saving by investigating it by considering it. According to the European Anti-Fraud Organization classification, the different typologies of corruption are based on the standard and financial flow characteristics involved in fraudulent processes [5][6].

B. Insurance Data Mining

Traditional methods for fraud detection are time-consuming and require complex investigations, making it challenging to deal with different knowledge domains. Machine learning algorithms have been diversely applied to predict future results from structured, unstructured, or complex data. The driving factor of these tools is the vast increase in data repositories at an enormous speed collected from various data sources. Various ML tools have evolved, reducing complexity and offering good facilities in various research domains, including Fraud detection, insurance, Transport engineering, Telecommunication, E-commerce, Web data analysis, Stock market, Crime and Terrorism, Aircraft maintenance, customs, computer intrusion, Geographical and Spatial data, Education, Social network analysis, Sports, and Software engineering, among others. The majority of these practices involve classification tasks that include the process of decision-making. The supervised model's primary principle is predicting the unknown outcome by training and learning with known or past data[7]

II. LITERATURE REVIEW

A literature review plays a significant role in gaining insight into the research problem. The literature review's primary function is to determine the work, theoretical and empirical, which has been done before so that it helps in the delineation of the plethora of problems in the area. The model discussed by Shin, Hyunjung et al. (2012) detects abusive utilization patterns in outpatient clinics claims data of Korean internal medicine clinics. The model is about an automated abuse system, and it is defined in two stages. The scoring stage is to quantify the degree of abusiveness in the billing patterns. In the stage of segmentation of the abnormal providers with the same utilization patterns [8].Christensen et al. (2018) objective of the work is to combine the natural language word embedding and network modeling for disease prediction in the asynchronous medical claims data. Most models use NN, LR, and deep learning models to predict future medical conditions. The work discussed embeddings' usefulness in capturing the temporal relationships between the observed diagnosis/procedures with the previous techniques. The non-deep learning XGBoost is used to demonstrate the cases with superior results. Their work's limitation is combining more models to obtain better outcomes [9].Matthew Herland et al. (2018) developed an anti-fraud healthcare system for multiple Medicare data. The logistic regression method yielded the best results for the part B dataset. The results generated based on the hypothesis and which part of Medicare data will commit fraud cannot be identified. The work concluded by finding fraud behaviors in submitted payments [10]. He (2019) extends to severe class imbalance effects while removing the fraud instances to determine the class rarity on fraud detection performance. Performed experiments using RF under-sampling, GBT, and LR on both training and testing and assess the potential improvements in fraud detection on Part B and Part D and the

combined dataset. The AUC, SD, Min, and Max metrics were used to compare train and test data[11]. Matthew Herland et al. (2020) extend their work toward identifying medical fraud in provider claims data using data mining strategies. The two baseline algorithms used are Multinominal Naïve Bayes, and Logistic Regression are used to investigate the scam on U.S Medicare part B dataset. Compared to their previous work, the accuracy of prediction increased. The prediction of the physician's specialty is based on the type and number of procedures performed if they could predict other than their specialty, they can be suspicious /anomalous behaviors. To produce mixed results by adding the class grouping, class removal strategies, and baseline algorithms [12].The scope of this paper is primarily on comparative analysis of the aggregated synthetic CMS claims data to classify the fraud and un-fraud claims using supervised machine learning algorithms. Most existing systems are static, related to the claim, without looking into the suspicious behavior patterns relevant to relationships and exchanges among entities. Controlling fraud is a unique challenge for more incredible economic countries. Computer-based technology has been adopted by the healthcare sector, which predominantly increased the electronic health data collected in various forms. Developing a model for identifying fraudulent claims involves training and testing phases of ML algorithms with metrics.

III. METHODS OF MACHINE LEARNING

In machine learning, a classifier's primary objective is to categorize new input data into a specific category or class. Features in the dataset are important for training the model to predict the class of new data. Classification methods in machine learning automate operations and improve efficiency, with two types of learners: lazy and eager. Supervised learning uses regression and classification techniques to predict numeric and continuous variables and categorize data into class labels. Unsupervised learning trains models based on previous data and predicts new data based on observations. Clustering is used to find groups in input data based on similarity, and dimensionality reduction eliminates features without losing importance. Anomaly detection identifies outliers, and neural networks process complex data using units, connections, and biases.

A. Random Forest classifier (RFC)

RFC is a classifier that creates multiple decision trees from a randomly selected subset of the training data. The final prediction of the outcome/decision is derived from the voting process of different decision trees. The underlying concept of RFC is the building block of Decision Trees (DT). That is the reason it is popularly known as ensemble operational. RFC is mighty comparatively with other supervised models, as a large number of relatively uncorrelated models will outperform for results when they operate in a constitutional way. The errors will be corrected iteratively by protecting trees from each other. The predictions made by the individual trees have a low correlation with each other, but these will produce better results than the random guessing of a model. The bagging method brings a diversity of models by replacing a random sample from the dataset, as it results in creating different trees. The process of feature randomness also forces even more variation amongst the trees and creates low correlation, leading to more diversification. Finally, the feature randomness and bagging will create uncorrelated multiple trees. An accurate prediction can be made with feature selection, as they should possess good predictive power [17].

B. Logistic Regression (LR)

LR is one type of statistical analysis for prediction, the estimation of probability is based on the relationship between one or more independent and dependent variables. The study helps find the likelihood of event occurrence and the choice made. Most of the analytical approach is useful for prediction, discovering uncovering patterns, and preferable for analyzing complex digital data. The logistic regression types differ as binary, multinomial, and ordinal based on the number of outcomes. The principal method involved in LR is the cost function which estimates the connection between the dependent and independent variables, i.e., the difference between actual and expected values. The regularization method reduces the cost function and avoids overfitting [18].

C. Support Vector Machine (SVM)

SVM is one algorithm used for regression and classification. SVM creates a best line or decision to segregate n-dimensional space into classes that could help correct the prediction of new data points or samples. Here the best decision boundary is known as the hyperplane of SVM, and SVM, extreme points/support vectors choose those. The hyperplane can be a single line or multiple lines used to classify samples. The distance between the extreme points of the hyperplane called as margin. The objective of SVM is to maximize the margin is termed an optimal hyperplane. The cost and gradient function helps to maximize the margin between the support vector points.

SVM finds the closest point from both classes and decides the class of the new point. SVM is categorized into two types Linear and Non-linear SVM. The data distribution is carried out based on separating the dataset by using and not using a single straight line [19].

D. Artificial Neural Networks (ANNs)

ANNs are a biologically inspired brain network. ANNs is the one capable of different types of learning. ANNs network is constructed with three layers: input, hidden, and output. Input layers generally accept inputs in different formats provided by the users/programmers. The hidden layer performs calculations to find the hidden features/patterns and results and is carried to the output layer. The computations involve weighted sum and bias at the input layer, and the transfer/activation function will decide whether the node should be fired or not. Different activation functions are used for other applications/tasks. Hyper-parameter is a constant parameter whose values are set for the learning process, including the network's learning rate. The learning rate helps adjust the errors of the observations for model convergence. The process is known as backpropagation. A higher learning rate reduces the training time, but it lowers the accuracy. There are feedforward and feedback ANNs based on the flow of information for pattern/recognition/classification [20].

IV. METHODOLOGY

The methodology here applies different classifiers, compares their results, and evaluates the best classifier. The general approach of a classifier method carries two levels: training and testing, where the algorithm has to access the values of both predictor and goal attributes. It represents classification knowledge. The knowledge is essential to find the relationship between the predictor and goal attribute (class label attribute), allowing class prediction. Predictions are made to the test set with the actual class attribute value at the testing level. The technique aims to improve the prediction accuracy while classifying the test set hidden in the training phase. The classification rules construct the knowledge (classification model). The rule-based approach recommends decision, sequential covering rule-induction algorithms, ANN (Artificial Neural Networks), and SVM (Support Vector Machine) [21].

The following steps are very much standard for any classifier

1. Initialize the classifier
2. Train the classifier with the technique
3. Predict the target variable
4. Evaluate the performance of the classifier with different metric

V. RESULTS AND DISCUSSIONS WITH METHODS

A. Data collection and preparation

The experimentation collects from the Center for Medicaid Services in the U.S Department of Health and Human Services. CMS has various sections as Part B, Part D, and DMEPOS. The empirical comparison of different models on the dataset to identify and evaluates the number of fraud claims [15].

B. Evaluation metrics

1. Predictive Accuracy:The training set uses accuracy as the metric for a classification system, whereas the test set does not use it for computation. The elements of the confusion matrix measure it. The confusion matrix consists of an nxn matrix,n is the number of classes for the problem, which holds the incorrect and correct categorizations by the model. For instance, the confusion matrix illustrates the number of truly and falsely classified samples per class in a binary classification problem. The confusion matrix illustrates the positive and negative classes as true positives (TP) and true negatives (TN) and examples of incorrect samples as positive and negatives as the false positives (FP) and false negatives (FN). The objective of the classification model is to attain high predictive accuracy in the test case; in the training case, the accuracy should be trivial. The algorithm records the class of each training instance to find the actual class by trivially searching the dataset.

2. Area Under Receiver Operating Characteristics (AUROC);It is an another measure to estimate predictive accuracy based on sensitivity, specificity, and classification quality. Here the sensitivity is defined as the actual positive rate [16]. The other metric, specificity, is defined as the true negative rate. ROC is used for predictive accuracy measures for unbalanced classes, in which the predictions for each class are made separately. ROC analysis is carried between true positive rate (sensitivity) against false positive rate (1-sensitivity) for classification models and several parameter values. The analysis plots a curve among the parameters. ROC is preferred over standard accuracy measures for classification [20][21].

C. Research Procedures

The data is of 173384 tuples, 22 attributes, and the partition section of data 3:1 to train and test the model. Feature processing and learning are needed for model convergence. Hyperparameter tuning to use to avoid overfitting and underfitting of the model and stabilizes the validation scores [15].Table 1 shows the accuracy and AUC values results for RF, SVM, LR, and ANN ML algorithms. Accuracy is the most crucial metric for assessing an algorithm's performance. It will simply realize with a set of samples of the target class in the dataset. The algorithm focuses on the majority class for the prediction. Figure 1. shows the accuracy of all the methods was at an average level. AUC is that it shows the class distribution, and thus it's likely to suffer from the dataset's imbalance.

TABLE I. CLASSIFIER PERFORMANCE WITH ACCURACY AND AUC

Model	Classifier Performance	
Metric	Accuracy	AUC
Logistic Regression	80.2	50.2
Support Vector Machines	82.3	75.8
Random Forest	82.3	75.2
Artificial Neural Networks	80.6	50.4

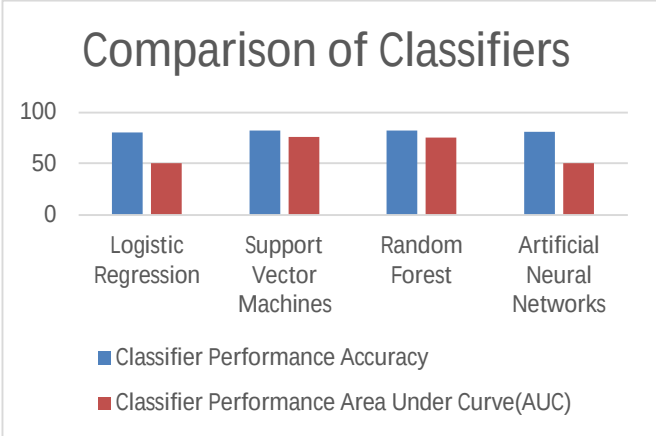

Figure 1. Performance of different classifiers

TABLE II. DIFFERENT STATISTICAL METRICS USED FOR ML METHODS TO EVALUATE THE PERFORMANCE AND THEIR DISCRIMINATION

Parameter		LR	SVM	RF	ANN
CV Score	Mean	0.50	0.50	0.500428	0.507336
	Stand. Dev.	0.0064	0.0064	0.006722998	0.0018613
	Min	0.487	0.489	0.4896601	0.4725037
	Max	0.5188	0.513	0.5133004	0.502521
Duration in seconds		37.189	39.4	39.445	48.322
Precision	0	0.17	0.17	0.17	0.80
	1	0.83	0.83	0.83	0.82
Recall	0	0.04	0.00	0.04	0.14
	1	0.96	0.76	0.96	0.90
F1-score	0	0.06	0	0.06	0.01
	1	0.89	0.91	0.89	0.67

The low values of AUC show the they do not produce reliable results, and they were primarily unable to predict all the target classes. The AUC values have noticeably increased for RF and SVM models, whereas they were not at a reasonable level for FR and ANNs models.

Table 2 shows the performance of ML methods on the CMS dataset with another set of metrics like CV score with mean, SD, min, max, duration of execution, precision, and recall. The results show that SVM, RF,ANNs models were at above-average levels shows the deviation of all the metrics.Table 3 shows the values of the classifier as to how the class distribution is done with the target class for the ML methods. The sensitivity and specificity of a model are described with the help of a confusion matrix. The useful matrix gives highly accurate results for both measures. The total distribution of 52,016 tuples for ML classifiers is shown in table 3 and referred to as TP, TN, FP, and FN.

TABLE III. CONFUSION MATRIX DESCRIPTION OF ML METHODS

Confusion Matrix	Number of instances	
Logistic Regression	399(TP)	8620(FP)
	1660(FN)	41337(TN)
Support Vector Machines	400(TP)	8019(FP)
	1600(FN)	41997(TN)
Random Forest	349(TP)	8670 (FP)
	1695(FN)	41302 (TN)
ANNs	454(TP)	8510(FP)
	1416(FN)	41636(TN)

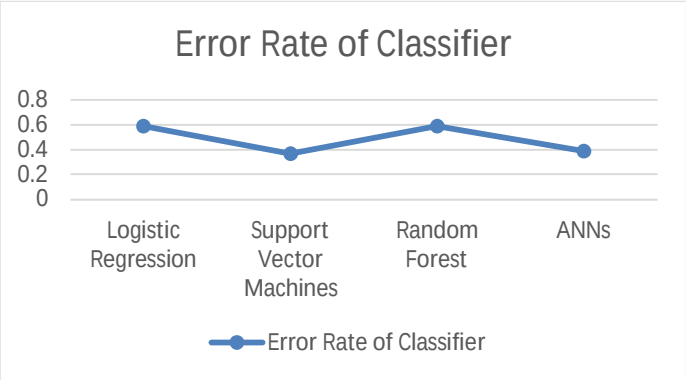

Figure 2. The error rate of ML classifiers

Finally, Figure 2. shows information about the error rate of classifiers and how the classifier picks samples that were used to design a model. In general, it tends to be biased optimistically. The true error rate of the classifier is invariably higher than the apparent error rate. The higher values and more prone to error at a higher level, and lower values are of good performers the paper evaluated FP+FN/Total no. of tuples dataset (TP+TN+FP+FN).

VI. CONCLUSION AND LIMITATIONS

Healthcare costs have risen due to the aging population, and health insurance policies have been developed to manage resources for all groups. A comparative analysis using different machine learning algorithms showed that SVM had superior accuracy in identifying fraudulent and non-fraudulent claims. The use of advanced machine learning algorithms can improve healthcare cost management, ensure efficient utilization of resources and identifying suspicious patterns. Feature selection, reduction, and transformation can help with model convergence, and resampling and sampling techniques can improve prediction accuracy by avoiding overfitting and underfitting. Adapting different sampling techniques to various ML methods can be useful for future work. This approach can help insurance companies choose the best model to enhance profits by avoiding fraudulent behavior. However, the limitation of this work is the need for real-time labeled data to develop an automated supervised model that can work in real-time and handle new incoming data. As the demand for healthcare data grows, there is a continuous need for fine-tuning data to improve the accuracy of models.

REFERENCES

[1] IAIS (2007) Report on the Survey on Preventing, Detecting and Remedying Fraud in Insurance, International Association of Insurance Supervisors, Basel.

[2] Morse, D., (2015) Tackling insurance fraud: law practice, www.converium.com/ 2500. asp.

[3] Hoyt, R.E., Mustard, D.B., and Powell, L.S., (2006) the effectiveness of state legislation in mitigating moral hazard: evidence from automobile insurance, Journal of Law & Economics, vol.49, No.2, 427-450.

[4] Canadian Coalition Against Insurance Fraud (2015) Insurance fraud information from Canada, available at: http://www.insurance-canada.ca/claims/canada/ CCAIF200110.php

[5] Naga Jyothi, P., Rajya Lakshmi, D., and Rama Rao, K.V.S.N., (2019) Performance on fraud detection in medical claims of healthcare data, IJITEE, vol.8, No.7, 1158-1165.

[6] Weiwei Lin,Ziming Wu,Longxin Lin,Angzhan Wen and Jin Li., An ensemble random forest algorithm for insurance big data analysis. IEEE Access Special section on Recent advances in computational intelligence paradigms for security and privacy fog and mobile edge computing, Vol 5, pp.16568-16575 (2017).

[7] Naga Jyothi, P., Rajya Lakshmi, D., and Rama Rao, K.V.S.N., (2022) Ph.D.,Thesis, Investigating and identifying fraudulent behaviors from multiple sources of medical claims data.

[8] Shin, H., Park, H., Lee, J. and Jhee, W.C., (2012) A scoring model to detect abusive billing patterns in health insurance claims, Expert Systems with Applications, vol.39, No.8, 7441-7450.

[9] Christensen, T., Frandsen, A., Glazier, S., Humpherys, J. and Kartchner, D.,(2018) Machine learning methods for disease prediction with claims data,IEEE International Conference on Healthcare Informatics (ICHI) 467-4674).

[10] Herland, M., Khoshgoftaar, T. M., & Bauder, R. A. (2018) Big data fraud detection using multiple medicare data sources. Journal of Big Data, vol.5, No.1, 1-21.

[11] Herland, M., Khoshgoftaar, T. M., & Bauder, R. A. (2019) The effects of class rarity on evaluation of supervised healthcare fraud detection models. Journal of Big Data, vol.6, No.1, 1-33.

[12] Herland, M., Khoshgoftaar, T. M., & Bauder, R. A. (2020) Approaches for identifying U.S. Medicare fraud in provider claims data, Health care management science,vol.23, No.2, 2-19.
[13] Naga Jyothi, P., Rajya Lakshmi, D., and Rama Rao, K.V.S.N., (2020) Identifying fraudulent behaviors in Healthcare Claims using Random Forest Classifier with SMOTEchnique, International Journal of e-Collaboration (IJeC), vol.16, No:4,30-47.
[14] Mitchell, T., (1997) Machine Learning and data mining, Communications of ACM, vol.42, No.11, 30-36.
[15] Witten, I.H., Frank, E., (2016) Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann.
[16] Caruana, R., Niculescu-Mizil, A., (2004) Data mining in metric space: an empirical analysis of supervised learning performance criteria. In: Proc. of the 10th Int. Conf. on Knowledge Discovery and Data Mining (KDD-04), ACM Press, 69–78.
[17] Witten, I.H., Frank, E., (2016) Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann.
[18] Gisele L.Pappa, Alex A, Freitas., (2009) Automating the design of data mining algorithms: An evolutionary computationapproach, Springer Science and Business Media.
[19] Pazzani, M.J., (2000) Knowledge discovery from data. IEEE Intelligent Systems and their applications, vol.15, No.2, 10-12.
[20] Caruana, R., Niculescu-Mizil, A., (2004) Data mining in metric space: an empirical analysis of supervised learning performance criteria. In: Proc. of the 10th Int. Conf. on Knowledge Discovery and Data Mining (KDD-04), ACM Press, 69–78.
[21] Flach, P., (2003) The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In: Proc. 20th Int. Conf. on Machine Learning (ICML-03), AAAI Press,194–201.