International Journal of Engineering Research and Science & Technology

IJERST

ISSN : 2319-5991

www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

https://doi.org/10.62643/ijerst.2025.v21.i2.pp408-419

Vol. 21, Issue 2, 2025

ADVANCING MEDICARE FRAUD DETECTION VIA MACHINE LEARNING AND SMOTE-ENN FOR IMBALANCED DATA

[1] *Shaik Abdullah, MCA Student, Department of MCA*
[2] *K Muddu Swamy, M.Tech, Assistant Professor, Department of MCA*
[12]*Dr KV Subba Reddy Institute of Technology, Dupadu, Kurnool*

ABSTRACT

The subject of healthcare fraud detection is always changing and has many obstacles to overcome, especially when dealing with skewed data. Prior research mostly concentrated on conventional machine learning (ML) methods, which often had trouble handling unbalanced data. This issue comes up in a number of ways. Random Oversampling (ROS) raises the danger of overfitting, the Synthetic Minority Oversampling Technique (SMOTE) introduces noise, and Random Undersampling (RUS) may result in the loss of important information. Furthermore, increasing assessment metrics, investigating hybrid resampling strategies, and optimising model performance are essential for reaching greater accuracy with unbalanced datasets. With an emphasis on the Medicare Part B dataset, we address the problem of unbalanced datasets in healthcare fraud detection in this study using a unique technique. The categorical feature "Provider Type" is first meticulously extracted from the dataset. This increases the variety within the minority class by enabling us to create new, synthetic instances by randomly reproducing preexisting kinds. Next, we use a hybrid resampling technique called SMOTE-ENN, which combines Edited Nearest Neighbours (ENN) with the Synthetic Minority Over-sampling Technique (SMOTE). By creating artificial samples and eliminating noisy data, this technique seeks to balance the dataset and increase the models' accuracy. We classify the examples using six machine learning (ML) models. We use standard measures like as accuracy, F1 score, recall, precision, and the AUC-ROC curve to assess performance. We emphasise how important the Area Under the Precision-Recall Curve (AUPRC) is for evaluating performance in situations with unbalanced datasets. With a score of 0.99 on all measures, the studies demonstrate that Decision Trees (DT) performed better than any classifier.

I. INTRODUCTION

Fraud is a major problem for healthcare systems across the world, affecting their morals and financial viability. One important component of the healthcare industry, the U.S. Medicare program, suffers significant financial losses as a result of these fraudulent activities. The Federal Bureau of Investigation estimates that 3–10% of healthcare expenses are attributable to healthcare fraud, which results in losses of $19 billion to $65 billion annually [1]. These illicit activities impact healthcare systems' operational effectiveness and credibility in addition to depleting financial resources. Because Medicare serves a wide and varied population, it is crucial to put robust and efficient fraud detection systems into place. Effective fraud detection is essential for safeguarding public money and ensuring equitable distribution of resources for patient care and essential healthcare services. The difficulty in detecting healthcare fraud stems from the dynamic and varied character of fraud schemes. In this dynamic context, traditional rule-based detection techniques are inadequate because they lack the flexibility and scalability needed to handle the complex nature of contemporary healthcare fraud. A subsection of artificial intelligence (AI), machine learning (ML), has shown remarkable effectiveness in detecting healthcare fraud, especially when processing the Medicare dataset that the US government releases

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp408-419
Vol. 21, Issue 2, 2025

every year [2]. For academics studying the detection of healthcare fraud, this dataset is an essential resource. This demonstrates the government's dedication to fighting fraud by providing experts with essential data, which makes it easier to create more complex ML-based fraud detection techniques. It is good at analysing big datasets to find abnormalities and fraud signs because of its capacity to learn from past data and adjust to new fraudulent tendencies. Because of its versatility, machine learning (ML) is a vital tool in the fight against healthcare fraud and may be used to develop responsive, effective systems for large-scale operations such as Medicare [3]. In this regard, however, machine learning (ML) shines. It can handle and analyse large datasets, identifying anomalies and patterns suggestive of fraud, thanks to its capacity to learn from past data and adapt to new fraudulent patterns. This feature makes machine learning (ML) an essential tool for developing more responsive and efficient fraud detection systems, particularly for large-scale programs like Medicare. It is a vital tool in the continuous battle against healthcare fraud because of its dynamic approach [3].

Using the Medicare dataset, recent research like those by [4], [5], [6], [7], and [8] show how ML approaches may be successfully used to discover fraudulent activity. There is a clear class disparity in the Medicare databases [9], which are published by the Centres for Medicare and Medicaid Services, with a disproportionate number of non-fraudulent cases compared to fraudulent cases. The effectiveness of machine learning algorithms used for fraud detection is severely hampered by this class disparity. A higher frequency of false negatives results from ML models' propensity to bias towards the majority class, in this instance, non-fraudulent transactions. As a direct result of the skewed training data, this phenomenon happens when the algorithm incorrectly classifies fraudulent activity as real [2], [10]. This dataset imbalance leads to the creation of machine learning models that perform less well than ideal when it comes to accurately identifying fraudulent activity. This shortcoming seriously compromises the overall efficacy and dependability of the fraud detection system in the healthcare industry. In order to address this issue, balanced datasets must be created so that machine learning algorithms can better identify the minority class, which in this case is fraudulent transactions. The algorithm's ability to identify subtle trends and anomalies suggestive of fraudulent activity depends on a balanced dataset [5].

The lack of attention paid to the problems caused by unbalanced data is a significant weakness in the research being done on healthcare fraud detection. The majority of research has focused on classification problems, paying little attention to the complex problem of data imbalance. Some researchers have started to fill this vacuum by using resampling approaches, despite the fact that there has been a noticeable lack of attention to data imbalances in healthcare fraud detection. These techniques include Synthetic Minority Over-sampling Technique (SMOTE) [12], Adaptive Synthetic Sampling Approach (ADASYN) [11], and Random Oversampling (ROS) [5]. To create a balanced dataset, Random Undersampling (RUS) is used concurrently to undersample the majority class [13]. Notwithstanding the effectiveness of these methods, problems still exist. For example, ROS approaches could be prone to overfitting, which might jeopardise the model's generalisability. In the meanwhile, there is a chance that using SMOTE will introduce noise into the dataset. Furthermore, there are a number of issues with RUS implementation, chief among them being the possibility of losing key data due to the discarding of critical information. The difficulty of resolving the class imbalance in healthcare fraud detection datasets is highlighted by the extensive trade-offs and considerations involved in each resampling strategy. We concentrate on three key areas to remedy the shortcomings noted in earlier studies:

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp408-419
Vol. 21, Issue 2, 2025

• Progressing studies on methods for handling unbalanced datasets
• Assessing resampling techniques, focussing on the limitations of SMOTE, which may introduce noise into the dataset, and ROS, which can result in overfitting.
• Analysing how RUS may affect the possible loss of crucial data, which can cause important fraud indications to be missed.

With a special emphasis on the Medicare Part B dataset, this work presents a unique method for addressing unbalanced datasets in healthcare fraud detection. The careful separation of the numerical and categorical properties is a significant advance that allows for the random creation of synthetic cases to enhance minority class diversity. By concurrently rebalancing the dataset and removing noisy data, our suggested Synthetic Minority Over-sampling technique with Edited Nearest Neighbours (SMOTE-ENN) hybrid resampling approach makes a considerable contribution. The dataset is then assessed using a variety of ensemble classifiers. To the best of our knowledge, this study suggests a method that combines the SMOTE-ENN methodology, a range of ensemble learning classifiers, and the independent production of categorical features. To further improve the thoroughness and resilience of our study, we also use the Area Under the Precision-Recall Curve (AUPRC) measure for assessment.

The following is a summary of this paper's primary contributions:

• Using the dataset's preexisting categories, construct the category feature "Provider. Type" at random.
• The SMOTE-ENN hybrid resampling technique is used to eliminate noisy data and balance the dataset.
• Using ensemble learning classifiers, the efficacy of the suggested strategy is assessed.
• Using the Area Under the Precision-Recall Curve (AUPRC) measure to assess model performance more effectively while dealing with an unbalanced dataset

This paper's structure is set up as follows: An review of the relevant literature is given in Section II, with a focus on research that made use of data balancing and machine learning. Section IV provides specifics on the suggested system. Section V presents the experimental data and a commentary. Section VI, which summarises the key findings, brings the article to a close.

II. LITERATURE SURVEY

"Healthcare fraud prevention: A crucial part of any cost-containment plan,"
Morris, L.
Fraudsters are attacking federal health care systems, such as Medicare and Medicaid, by lying to the government and taking advantage of its programs to embezzle public funds. It is impossible to quantify the entire scope of health care fraud. Nonetheless, according to the Federal Bureau of Investigation (FBI), in fiscal year 2009, fraudulent billings to public and commercial health care programs accounted for 3–10% of all health expenditures, or $75–$250 billion. Aggressive, creative, and persistent efforts are needed to prevent such abuses without placing an undue burden on authorised providers in order to safeguard taxpayers and beneficiaries.

"Explainable machine learning models for the detection of Medicare fraud,"
T. M. Khoshgoftaar, H. Wang, R. A. Bauder, and J. T. Hancock,
We use a unique ensemble supervised feature selection method to construct explainable machine learning models for Big Data. The method is used using publicly accessible insurance claims data from Medicare, the public health insurance program in the United States. By classifying extremely unbalanced Big Data, we tackle Medicare insurance fraud detection as a supervised machine learning job of anomaly detection. Enhancing model training efficiency and creating more comprehensible machine

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp408-419
Vol. 21, Issue 2, 2025

learning models for fraud detection are our goals for feature selection. We illustrate how our feature selection method lowers the dimensionality of the datasets by around 87.5% without sacrificing performance using two Big Data datasets that were obtained from two distinct sources of insurance claims data. Furthermore, machine learning models that have fewer dimensionality are less likely to overfit and are simpler to understand. As a result, our fundamental contribution—the description of our unique feature selection method—leads to an additional contribution to the field of automated Medicare insurance fraud detection applications. In order to explain our fraud detection models in terms of the meanings of the chosen features, we use our feature selection approach. We describe an ensemble supervised feature selection method that can be used to any set of machine learning algorithms that keep track of a list of feature priority values. As a result, researchers may readily use variants of the method we describe.

"Applying machine learning to healthcare opportunities and challenges,"
Alanazi, A.
The use of machine learning (ML) in healthcare has drawn a lot of interest. Big data and increased processing power provide a chance to use machine learning algorithms to benefit healthcare. Using techniques like logistic or linear regression, support vector machines, decision trees, LASSO regression, K Nearest Neighbour, and the Naive Bayes classifier, supervised learning is the kind of machine learning that may be used to forecast labelled data. Unsupervised machine learning algorithms are able to spot patterns in datasets without outcome information. These models may be used to anomaly or fraud detection. Several clinical decision support systems are examples of clinical uses of machine learning. The identification and prediction of groups at high risk for certain bad health outcomes, as well as the creation of public health interventions aimed at these populations, constitute a significant public health application of machine learning. For medical professionals to effectively direct and interpret research in this field, a variety of ML-related concepts must be incorporated into the curriculum.

"Medicare fraud detection through machine learning techniques with excluded provider labels,"
T. M. Khoshgoftaar and R. A. Bauder,
There are more essential medical demands and expenses associated with the general rise in the senior population. Medicare is a healthcare program in the United States that offers insurance to those 65 and older, mostly to help them with the cost of medical treatment. Nevertheless, the expense of healthcare is enormous and still rising. One of the main causes of these rising healthcare costs is fraud. In order to identify fraudulent Medicare providers, our research offers a thorough analysis using machine learning techniques. We construct and evaluate three distinct learners using publicly accessible Medicare data and provider exclusions for fraud labelling. Given the limited number of real fraud labels, we use random undersampling to create four class distributions in order to mitigate the effects of class imbalance. With average AUC scores of 0.883 and 0.882, respectively, and low false negative rates, our findings demonstrate that the C4.5 decision tree and logistic regression learners perform the best in terms of fraud detection, especially for the 80:20 class distribution. We effectively show that using machine learning with random undersampling to identify Medicare fraud is effective.

"Detecting Medicare fraud through machine learning techniques,"
T. M. Khoshgoftaar and R. A. Bauder,
Affordable healthcare is essential to people's lives, particularly for the growing senior population. One such health care program is Medicare. Although claims fraud is a significant cause of rising healthcare expenses, fraud

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp408-419
Vol. 21, Issue 2, 2025

identification may mitigate its effects. In this work, we evaluate a number of machine learning techniques for Medicare fraud detection. Using four performance indicators and reducing class imbalance by oversampling and an 80-20 undersampling technique, we compare supervised, unsupervised, and hybrid machine learning techniques. Using fraud labels from the List of Excluded Individuals/Entities database, we classify the 2015 Medicare data according to provider categories. Our findings demonstrate that it is feasible to successfully identify fake providers, with the 80-20 sample strategy showing the greatest performance among learners. Additionally, supervised approaches outperformed unsupervised or hybrid approaches; however, the findings differed according on the kind of provider and the class imbalance sampling methodology.

"Developing prediction models and identifying critical elements of health insurance fraud through machine learning techniques,"

According to V. Nalluri, J.-R. Chang, L.-S. Chen, and J.-C. Chen, 3–10% of annual medical expenses are attributable to health insurance fraud. Allowing the expansion of fraudulent activity will have permanent effects on the healthcare system. However, the volume and complexity of medical data make it challenging to analyse using conventional statistical techniques. As a result, one of the key answers is now machine learning algorithms. One important concern is whether the learning approach can remain stable and provide a better suitable response when presented with diverse data. Few research seek to identify the key components of medical fraud and choose the best machine learning technique, despite the fact that many related studies concentrated on medical insurance fraud and evaluation. In order to identify the most effective machine learning technique for detecting medical fraud, this study employed four machine learning techniques—Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), and Multilayer Perceptrons (MLP)—as well as two unpublished datasets that may reveal new information. We also identified 19 key elements of medical insurance fraud from the DT data and categorised them into four groups: beneficiary, applicable insurance claims amount, medical service providers, and Healthcare Common Procedure Coding System (HCPCS). Experiments' findings may provide insurance management useful recommendations for setting up an automated audit system to stop medical fraud.

III. SYSTEM ANALYSIS AND DESIGN

EXISTING SYSTEM

Diverse and creative methods for identifying healthcare fraud have been made possible by recent developments in AI, particularly machine learning. By using the large Part D Medicare dataset, which contains over 175 million records, the authors in [16] sought to enhance decision-tree-based ensemble approaches for healthcare fraud detection. With an emphasis on business heuristics, provider-prescriber interactions, and client demographics, the authors of [17] presented an ML system that converts prescription claims into statistical modelling characteristics. In order to identify Medicare fraud, the research by [2] used an ensemble feature selection strategy in ML models. This method decreased data complexity and increased explainability. In the study suggested by [18], Texas Medicaid prescription claims were preprocessed and feature engineered in order to establish a Bayesian Belief Network (BBN) model for healthcare fraud detection. In terms of interpretability and scalability, this method fared better than baseline models.

The authors of [19] focused on using a data-centric AI strategy to identify Medicare fraud in the United States. With careful feature engineering and data preparation, this greatly improved the performance of ML models. Their method performed better on Medicare fraud classification tests than standard datasets. Reference [6] suggested a research that uses four

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp408-419
Vol. 21, Issue 2, 2025

machine learning algorithms to identify cases of healthcare fraud. They categorised the 19 key characteristics they found in their investigation into four main groups. Examining the papers, we can see that a variety of strategies, including ensemble methods, decision-tree-based techniques, and BBN, are used to identify fraud. Furthermore, a number of studies stress how crucial feature engineering, feature selection, and data preparation are to improving the model's performance. One prevalent drawback, nevertheless, is the experimentation's dependence on the highly unbalanced Medicare dataset; this problem is still mostly unresolved and may lead to erroneous results.

In order to address the issue of unbalanced data, the article [20] experimented with several class distributions in their machine learning models. The authors addressed the data imbalance by applying six machine learning models over seven class distributions using the Medicare Part B dataset. According to the findings, using a 90:10 ratio of fraud instances to non-fraud cases fared better than other approaches. The authors of the research [21] used six sampling strategies to balance the Medicare dataset using machine learning models for classification in order to solve the issue of the dataset's imbalance. The results of the investigation showed that RUS consistently produced good outcomes for all ML models. In [22], a semantic embedding method was suggested.

In order to transform healthcare procedure codes (HCPCS) from the Medicare fraud dataset into semantic embeddings, the author suggested a semantic embedding technique. A straightforward undersampling technique was used in the study to overcome the unbalanced data problem. In [23], a further semantic embedding method was put forward. Using pre-trained (Global Vectors for Word Representation (GloVe), Medical Word2Vec (Med-W2V)) and bespoke (HcpcsVec, RxVec) embeddings using Medicare claims data, the authors created semantic embeddings for several kinds of medical providers. Several machine learning methods were used to verify this approach, which enhanced the depiction of provider specialities. Additionally, the research used random over-sampling (ROS) and under-sampling strategies to address the problem of unbalanced data.

The authors of the research [24] suggested using unsupervised DL approaches to identify medical claims that overuse procedure codes. The test set was made up of outliers that could be fraudulent cases in order to address the unbalanced data. The performance of ML classifiers in the Medicare unbalanced dataset was the main focus of the study [25]. To solve class imbalances, the authors used a variety of ensemble learning strategies with the RUS approach. In a different study, [26], the RUS technique was used to overcome the imbalance problem and investigate the categorisation of healthcare fraud using the Medicare dataset, which is very unbalanced. According to the findings, RUS reduced the quantity of the training data while increasing the AUC values. The authors of the research [11] suggested using two data balancing methods: ADASYN and the Class Weighing Scheme (CWS). Additionally, the authors used a variety of machine learning methods to categorise occurrences.

Disadvantages

• Progressing studies on methods for handling unbalanced datasets

• Assessing resampling techniques, paying particular attention to the limitations of ROS, which may lead to overfitting, and SMOTE, which might introduce noise into the dataset.

• Analysing how RUS may affect the possible loss of crucial data, which may cause important fraud signs to be missed.

PROPOSED SYSTEM

With a special emphasis on the Medicare Part B dataset, this work presents a unique method for addressing unbalanced datasets in healthcare fraud detection. The careful separation of the

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

https://doi.org/10.62643/ijerst.2025.v21.i2.pp408-419

Vol. 21, Issue 2, 2025

numerical and categorical properties is a significant advance that allows for the random creation of synthetic cases to enhance minority class diversity. By concurrently rebalancing the dataset and removing noisy data, our suggested Synthetic Minority Over-sampling technique with Edited Nearest Neighbours (SMOTE-ENN) hybrid resampling approach makes a considerable contribution. The dataset is then assessed using a variety of ensemble classifiers. To the best of our knowledge, this study suggests a method that combines the SMOTE-ENN methodology, a range of ensemble learning classifiers, and the independent production of categorical features. Furthermore, we use the Area Under the Precision-Recall Curve (AUPRC) measure for assessment, which improves the thoroughness and resilience of our study.

Advantages

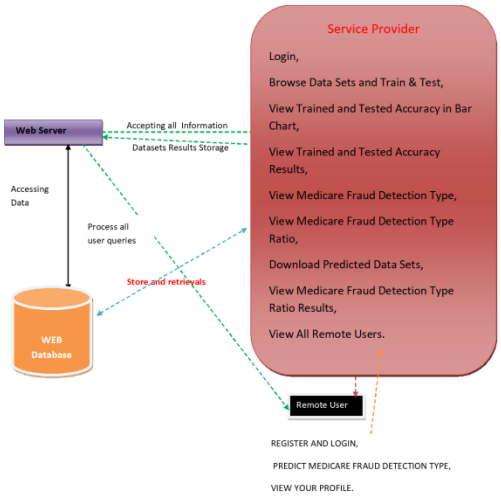
• Using the dataset's preexisting categories, construct the category feature "Provider. Type" at random.

• The SMOTE-ENN hybrid resampling technique is used to eliminate noisy data and balance the dataset.

• Using ensemble learning classifiers, the efficacy of the suggested strategy is assessed.

• Using the Area Under the Precision-Recall Curve (AUPRC) measure to assess model performance more effectively while dealing with an unbalanced dataset.

SYSTEM ARCHITECTURE



IV. IMPLEMENTATION

Modules

Service Provider

The Service Provider must use a working user name and password to log in to this module. He may do many tasks after successfully logging in, including Train & Test Data Sets, See the Accuracy of Trained and Tested Datasets in a Bar Chart View Accuracy Results for Trained and Tested Datasets, Download Predicted Data Sets, View Cyber Attack Prediction Status Ratio, and View Cyber Attack Prediction Status See the results of the Cyber Attack Prediction Status Ratio. See Every Remote User.

View and Authorize Users

The administrator may see a list of all registered users in this module. Here, the administrator may see the user's information, like name, email, and address, and they can also grant the user permissions.

Remote User

A total of n users are present in this module. Before beginning any actions, the user needs register. Following registration, the user's information will be entered into the database. Following a successful registration, he must use his password and authorised user name to log in.

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp408-419
Vol. 21, Issue 2, 2025

Following a successful login, the user may do tasks including registering and logging in, predicting the status of cyberattacks, and seeing their profile.

ALGORITHM

Logistic regression Classifiers

The relationship between a collection of independent (explanatory) factors and a categorical dependent variable is examined using logistic regression analysis. When the dependent variable simply has two values, like 0 and 1 or Yes and No, the term logistic regression is used. When the dependent variable contains three or more distinct values, such as married, single, divorced, or widowed, the technique is sometimes referred to as multinomial logistic regression. While the dependent variable's data type differs from multiple regression's, the procedure's practical application is comparable.

When it comes to categorical-response variable analysis, logistic regression and discriminant analysis are competitors. Compared to discriminant analysis, many statisticians believe that logistic regression is more flexible and appropriate for modelling the majority of scenarios. This is due to the fact that, unlike discriminant analysis, logistic regression does not presume that the independent variables are regularly distributed.

Both binary and multinomial logistic regression are calculated by this software for both category and numerical independent variables. Along with the regression equation, it provides information on likelihood, deviance, odds ratios, confidence limits, and quality of fit. It does a thorough residual analysis that includes diagnostic residual plots and reports. In order to find the optimal regression model with the fewest independent variables, it might conduct an independent variable subset selection search. It offers ROC curves and confidence intervals on expected values to assist in identifying the optimal classification cutoff point. By automatically identifying rows that are not utilised throughout the study, it enables you to confirm your findings.

Naïve Bayes

The supervised learning technique known as the "naive bayes approach" is predicated on the straightforward premise that the existence or lack of a certain class characteristic has no bearing on the existence or nonexistence of any other feature. However, it seems sturdy and effective in spite of this. It performs similarly to other methods of guided learning. Numerous explanations have been put forward in the literature. We emphasise a representation bias-based explanation in this lesson. Along with logistic regression, linear discriminant analysis, and linear SVM (support vector machine), the naive bayes classifier is a linear classifier. The technique used to estimate the classifier's parameters (the learning bias) makes a difference.

Although the Naive Bayes classifier is commonly used in research, practitioners who want to get findings that are useful do not utilise it as often. On the one hand, the researchers discovered that it is very simple to build and apply, that estimating its parameters is simple, that learning occurs quickly even on extremely big datasets, and that, when compared to other methods, its accuracy is rather excellent. The end users, however, do not comprehend the value of such a strategy and do not get a model that is simple to read and implement.

As a consequence, we display the learning process's outcomes in a fresh way. Both the deployment and comprehension of the classifier are simplified. We discuss several theoretical facets of the naive bayes classifier in the first section of this lesson. Next, we use Tanagra to

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp408-419
Vol. 21, Issue 2, 2025

apply the method on a dataset. We contrast the outcomes (the model's parameters) with those from other linear techniques including logistic regression, linear discriminant analysis, and linear support vector machines. We see that the outcomes are quite reliable. This helps to explain why the strategy performs well when compared to others. We employ a variety of tools (Weka 3.6.0, R 2.9.2, Knime 2.1.1, Orange 2.0b, and RapidMiner 4.6.0) on the same dataset in the second section. Above all, we make an effort to comprehend the outcomes.

Random Forest

Random forests, also known as random decision forests, are ensemble learning techniques that build a large number of decision trees during training for tasks like regression and classification. The class chosen by the majority of trees is the random forest's output for classification problems. The mean or average forecast of each individual tree is given back for regression tasks. The tendency of decision trees to overfit to their training set is compensated for by random decision forests. Although random forests are less accurate than gradient enhanced trees, they often perform better than choice trees. However, their performance may be impacted by data peculiarities.

Tin Kam Ho[1] developed the first algorithm for random decision forests in 1995 by using the random subspace technique, which in Ho's definition is a means of putting Eugene Kleinberg's "stochastic discrimination" approach to classification into practice.

Leo Breiman and Adele Cutler created an algorithm extension and filed for a trademark in 2006 for "Random Forests" (owned by Minitab, Inc. as of 2019). The extension builds a set of decision trees with controlled variance by combining Breiman's "bagging" concept with random feature selection, which was initially proposed by Ho[1] and then separately by Amit and Geman[13].

Businesses often employ random forests as "blackbox" models since they need minimal setup and provide accurate forecasts across a variety of inputs.

SVM

The goal of a discriminant machine learning approach in classification problems is to identify a discriminant function that can accurately predict labels for newly acquired instances based on an independent and identically distributed (iid) training dataset. A discriminant classification function takes a data point x and assigns it to one of the several classes that are part of the classification job, in contrast to generative machine learning techniques that call for calculations of conditional probability distributions. Discriminant techniques are less effective than generative approaches, which are mostly used when prediction entails the identification of outliers. However, they need less training data and processing resources, particularly when dealing with a multidimensional feature space and when just posterior probabilities are required. Finding the equation for a multidimensional surface that optimally divides the various classes in the feature space is the geometric equivalent of learning a classifier.

SVM is a discriminant approach that, unlike genetic algorithms (GAs) or perceptrons, which are both often used for classification in machine learning, always returns the same optimum hyperplane value since it solves the convex optimisation issue analytically. The initialisation and termination criteria have a significant impact on the solutions for perceptrons. While the perceptron and GA classifier models are distinct every time training is started, training yields uniquely specified SVM model parameters for a given training set for a certain kernel that converts

International Journal of Engineering Research and Science & Technology

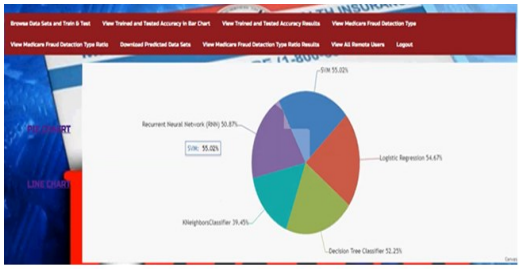
ISSN 2319-5991 www.ijerst.com

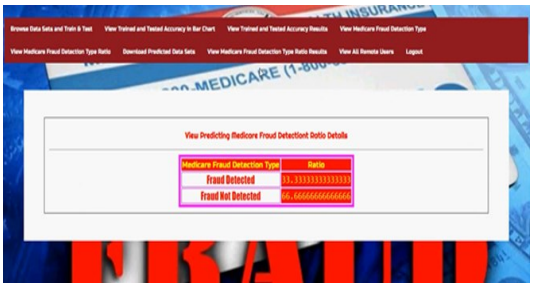
https://doi.org/10.62643/ijerst.2025.v21.i2.pp408-419

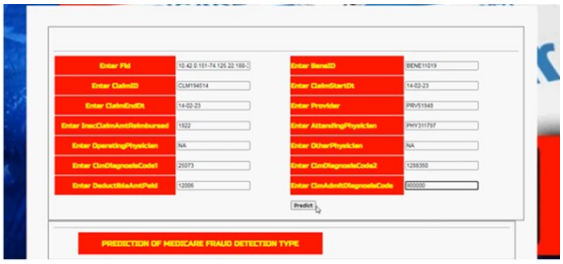
Vol. 21, Issue 2, 2025

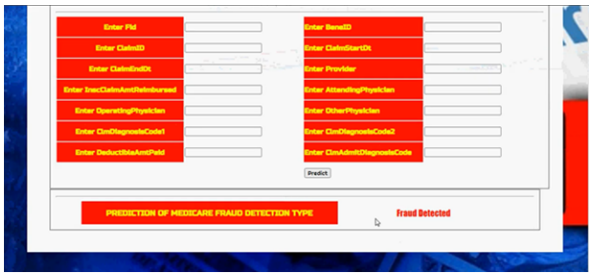
the data from the input space to the feature space. The only goal of GAs and perceptrons is to reduce training error, which will result in several hyperplanes satisfying this criterion.

V. SCREEN SHOTS









VI. CONCLUSION

By presenting a unique machine learning framework based on the SMOTE-ENN hybrid resampling approach, this research highlights the need of addressing unbalanced data in healthcare fraud detection. By removing noisy data and producing synthetic samples, this technique efficiently balances datasets and improves the accuracy of the model. The use of the AUC and AUPRC as assessment measures is another facet of our research. A comprehensive examination of the models' performance was made easier by these measures, with the AUPRC showing particular importance when dealing with unbalanced datasets. As a result, this method provides a foundation for future researchers to use in their efforts to identify healthcare fraud. Future studies will examine how well SMOTE-ENN performs in various healthcare fraud situations and combine it with cutting-edge AI

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp408-419
Vol. 21, Issue 2, 2025

tools like deep learning (DL) to improve the efficacy of fraud detection techniques.

REFERENCES

[1] L. Morris, ‘‘Combating fraud in health care: An essential component of any cost containment strategy,’’ Health Affairs, vol. 28, no. 5, pp. 1351–1356,Sep. 2009.

[2] J. T. Hancock, R. A. Bauder, H. Wang, and T. M. Khoshgoftaar, ‘‘Explainable machine learning models for medicare fraud detection,’’ J.Big Data, vol. 10, no. 1, p. 154, Oct. 2023.

[3] A. Alanazi, ‘‘Using machine learning for healthcare challenges and opportunities,’’ Informat. Med. Unlocked, vol. 30, 2022, Art. no. 100924.

[4] R. A. Bauder and T. M. Khoshgoftaar, ‘‘The detection of medicare fraud using machine learning methods with excluded provider labels,’’ in Proc. Thirty-First Int. Flairs Conf., 2018, pp. 1–6.

[5] R. A. Bauder and T. M. Khoshgoftaar, ‘‘Medicare fraud detectionusing machine learning methods,’’ in Proc. 16th IEEE Int. Conf.Mach. Learn. Appl. (ICMLA), Dec. 2017, pp. 858–865. [Online]. Available:http://ieeexplore.ieee.org/document/8260744/

[6] V. Nalluri, J.-R. Chang, L.-S. Chen, and J.-C. Chen, ‘‘Building prediction models and discovering important factors of health insurance fraud using machine learning methods,’’ J. Ambient Intell. Humanized Comput.,vol. 14, no. 7, pp. 9607–9619, Jul. 2023.

[7] P. Dua and S. Bais, ‘‘Supervised learning methods for fraud detection in healthcare insurance,’’ in Machine Learning in Healthcare Informatics (Intelligent Systems Reference Library), vol. 56, S. Dua, U. Acharya, and P. Dua, Eds. Berlin, Germany: Springer, 2014, doi: 10.1007/978-3-642-40017-9_12.

[8] R. Bauder, R. da Rosa, and T. Khoshgoftaar, ‘‘Identifying medicare provider fraud with unsupervised machine learning,’’ in Proc. IEEE Int. Conf. Inf. Reuse Integr. (IRI), Jul. 2018, pp. 285–292.

[9] Centers for Medicare and Medicaid Services. (2017). Research, Statistics, Data, and Systems. [Online]. Available: https://www.cms.gov/researchstatistics-data-and-systems/research-statistics-data-and-systems.html

[10] P. Brennan, ‘‘A comprehensive survey of methods for overcoming the class imbalance problem in fraud detection,’’ Inst. Technol. Blanchardstown Dublin, Dublin, Ireland, Tech. Rep., 2012.

[11] N. Agrawal and S. Panigrahi, ‘‘A comparative analysis of fraud detection in healthcare using data balancing & machine learning techniques,’’ in Proc. Int. Conf. Commun., Circuits, Syst. (IC3S), May 2023, pp. 1–4.

[12] M. Herland, R. A. Bauder, and T. M. Khoshgoftaar, ‘‘The effects of class rarity on the evaluation of supervised healthcare fraud detection models,’’ J. Big Data, vol. 6, no. 1, pp. 1–33, Dec. 2019.

[13] J. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, ‘‘The effects of random undersampling for big data medicare fraud detection,’’ in Proc. IEEE Int. Conf. Service-Oriented Syst. Eng. (SOSE), Aug. 2022,pp. 141–146.

[14] A. Mehbodniya, I. Alam, S. Pande, R. Neware, K. P. Rane, M. Shabaz, and M. V. Madhavan, ‘‘Financial fraud detection in healthcare using machine learning and deep learning techniques,’’ Secur. Commun. Netw., vol. 2021,pp. 1–8, Sep. 2021.

[15] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, ‘‘Learning from class-imbalanced data: Review of methods and applications,’’ Expert Syst. Appl., vol. 73, pp. 220–239, May 2017.

[16] J. Hancock and T. M. Khoshgoftaar, ‘‘Optimizing ensemble trees for big data healthcare fraud detection,’’ in Proc. IEEE 23rd Int. Conf. Inf. Reuse Integr. Data Sci. (IRI), Aug. 2022, pp. 243–249.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

https://doi.org/10.62643/ijerst.2025.v21.i2.pp408-419

Vol. 21, Issue 2, 2025

[17] N. Kumaraswamy, M. K. Markey, J. C. Barner, and K. Rascati, ‘‘Feature engineering to detect fraud using healthcare claims data,’’ Expert Syst. Appl., vol. 210, Dec. 2022, Art. no. 118433.

[18] N. Kumaraswamy, T. Ekin, C. Park, M. K. Markey, J. C. Barner, and K. Rascati, ‘‘Using a Bayesian belief network to detect healthcare fraud,’’ Expert Syst. Appl., vol. 238, Mar. 2024, Art. no. 122241.

[19] J. M. Johnson and T. M. Khoshgoftaar, ‘‘Data-centric AI for healthcare fraud detection,’’ Social Netw. Comput. Sci., vol. 4, no. 4, p. 389,May 2023.

[20] R. A. Bauder and T. M. Khoshgoftaar, ‘‘The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data,’’ Health Inf. Sci. Syst., vol. 6, no. 1, pp. 1–14, Dec. 2018.

[21] R. A. Bauder, T. M. Khoshgoftaar, and T. Hasanin, ‘‘Data sampling approaches with severely imbalanced big data for medicare fraud detection,’’ in Proc. IEEE 30th Int. Conf. Tools Artif. Intell. (ICTAI),Nov. 2018, pp. 137–142.

[22] J. M. Johnson and T. M. Khoshgoftaar, ‘‘Hcpcs2Vec: Healthcare procedure embeddings for medicare fraud prediction,’’ in Proc. IEEE 6th Int. Conf. Collaboration Internet Comput. (CIC), Dec. 2020, pp. 145–152.

[23] J. M. Johnson and T. M. Khoshgoftaar, ‘‘Medical provider embeddings for healthcare fraud detection,’’ Social Netw. Comput. Sci., vol. 2, no. 4, p. 276,Jul. 2021. [Online]. Available: https://link.springer.com/10.1007/s42979-021-00656-y

[24] M. Suesserman, S. Gorny, D. Lasaga, J. Helms, D. Olson, E. Bowen, and S. Bhattacharya, ‘‘Procedure code overutilization detection from healthcare claims using unsupervised deep learning methods,’’ BMC Med. Informat. Decis. Making, vol. 23, no. 1, p. 196, Sep. 2023.

[25] J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, ‘‘Evaluating classifier performance with highly imbalanced big data,’’ J. Big Data, vol. 10, no. 1, p. 42, Apr. 2023.