Indonesian Journal of Electrical Engineering and Computer Science
Vol. 32, No. 3, December 2023, pp. 1686~1694
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v32.i3.pp1686-1694

Ensemble learning based health care claim fraud detection in an imbalance data environment

Shweta S. Kaddi[1], Malini M. Patil[2]
[1]Department of Computer Science and Engineering, J.S.S. Academy of Technical Education,
Visvesvaraya Technological University (VTU), Bengaluru, India
[2]Department of Computer Science and Engineering, R V Institute of Technology and Management,
Visvesvaraya Technological University (VTU), Bengaluru, India

Article Info

Article history:

Received May 17, 2023
Revised Aug 14, 2023
Accepted Sep 25, 2023

Keywords:

AUC-ROC curve
Ensemble learning
Health care fraud detection
Logistic regression
SMOTE
Stack ensemble learning

ABSTRACT

Healthcare fraud has become a common encounter in the healthcare finance industry. The financial security of healthcare payers and providers is seriously impacted by healthcare fraud. When incorrect or exaggerated medical services are invoiced for reimbursement, fraudulent healthcare claims result. The effective operation of the healthcare system depends on the detection of such fraudulent actions. This paper develops a healthcare claim fraud detection method based on ensemble learning. Stack ensemble learning algorithm performance is compared to that of methods such as multi-layer perceptron (MLP) classifier, support vector classifier (SVC), logistic regression (LR), and decision tree (DT) algorithm. Because of the healthcare data imbalance, the normal transaction is significantly higher than the fraudulent transaction. The machine learning (ML) algorithm performs poorly because imbalanced data causes it to be biased toward the majority class. As a result, the data is unsampled using the synthetic minority oversampling technique (SMOTE) technique to provide balanced data. The experimental results show that for the identification of healthcare claim fraud, the ensemble learning strategy greatly outperforms single learning algorithms. The stack ensemble learning outperforms all the area under the curve for the receiver-operating characteristic (AUC ROC) curves from various algorithms, and the AUC-ROC curve is determined to be producing results that are adequate for all approaches.

Corresponding Author:

Shweta S. Kaddi
Department of Computer Science and Engineering, J.S.S. Academy of Technical Education
Visvesvaraya Technological University (VTU)
Belagavi, Karnataka, India
Email: shwetakaddi@gmail.com

1. INTRODUCTION

Healthcare insurance providers' and policyholders' capacity to maintain a healthy financial position is seriously impacted by healthcare fraud. When misleading or inflated medical services are billed for reimbursement, healthcare claims are fraudulent. The effective operation of the healthcare system depends on the detection of such fraudulent actions. However, because of the volume and complexity of the data involved, identifying healthcare fraud is a difficult undertaking. Traditional fraud detection techniques, such as rule-based algorithms and outlier identification, are ineffective in correctly identifying fraudulent claims. To address this issue, machine learning (ML)-based approaches have been proposed to detect healthcare fraud. These approaches use algorithms to learn patterns from the data and predict the likelihood of a claim being fraudulent. However, using a single learning algorithm may not provide accurate results due to the complexity and diversity of the data.

Traditional methods of fraud detection are often inadequate, as they rely on rule-based systems that are unable to adapt to new forms of fraud. Instead, ML algorithms like deep learning algorithms, capable of analyzing large amounts of data and identifying complex patterns that may indicate fraudulent activity is implemented [1]. Convolutional neural network (CNN) is used as the deep learning method for medical fraud detection in the classification solution developed. A real-world healthcare claim dataset is applied on the trained model and observed that the accuracy performance is high for the deep learning method employed. The sequence of events is subjected to sequence mining algorithms to find recurring patterns of transactions that might point to fraud in healthcare claims [2]. The findings demonstrated that in terms of detection accuracy, precision, and recall, the suggested architecture performed better than both the conventional rule-based approach and the ML-based approach. On the health dataset, the Apriori method is used to extract association rules from the data. A behavior model that can be used to identify abnormal behavior was built using the rules that emerged from the experiment. The behavior model was evaluated using a dataset of 2,000 patient records, and it was 91.5% accurate at detecting abnormal behavior [3]. To find anomalous behavior in patient data, manifold learning, and outlier detection algorithms are developed. To find outliers in the data, local outlier factor (LOF) technique is used. The ensuing outliers were considered to be a sign of fraudulent activity [4]. The suggested method analyses insurance claim data and spots possible fraudulent claims by combining rule-based and machine-learning techniques. The results show that the unique method of using natural language processing to extract clinical concepts is effective [5]. In many real-world applications, imbalanced datasets are widespread, where the proportion of instances in one class is much higher than the other (s). Due to their propensity to be biased in favor of the majority class, ML models can perform poorly in the presence of this imbalance due to their low classification accuracy for the minority class. Thus, resampling methods are used to obtain the balance between both the higher and lower instances [6]. The unbalanced dataset was initially preprocessed before training individual classifiers using a conventional classification methodology. The difficulty of classifying each sample was then determined using a distance-based measure, and weights were given to each classifier depending on how well they performed on the challenging examples. Without compromising performance on the sample data, the resulting ensemble model was able to increase classification performance on the challenging examples [7].

The relevant features were extracted using a feature extraction technique after the dataset had been preprocessed. They then trained an ensemble of classifiers using various classification algorithms, such as support vector machines (SVM), decision trees (DTs), and random forests (RFs), in order to balance the class distribution. While retaining the performance on the majority classes, the resulting ensemble model was able to increase classification performance for the minority classes [8]. Prior to using a deep learning neural network to extract the pertinent features, the data is preprocessed. They then used several methods, including DTs, RFs, and SVM, to train a number of models. To increase the precision of the predictions, the generated models were subsequently integrated using a stacking ensemble learning approach [9]. Overall, employing deep learning neural networks and stacking ensemble learning, the paper proposes an intriguing method for forecasting the CSI 300 index. The results show the usefulness of a novel strategy that uses numerous models and a stacking ensemble approach [9]. Soil moisture data is gathered from several sources and then trained support vector regression, RF regression, and gradient-boosting regression models using this data. Using a stacked generalization strategy, which entails training a meta-model on the predictions of the basic models, they then merged the predictions of these models. The final soil moisture forecasts are then based on the developed meta-model [10]. Discussion is held regarding the synthetic minority oversampling technique (SMOTE) technique and its variants, such as borderline-SMOTE, safe-level SMOTE, and adaptive synthetic sampling (ADASYN). SMOTE's drawbacks and suggestion for a number of ways to enhance it is discussed. These include integrating SMOTE with other approaches, modifying SMOTE's parameters, and employing other data creation techniques [11]. To overcome the issue of unbalanced sentiment analysis, SMOTE is used. The effectiveness of various feature representation techniques, such as bag-of-words, n-grams, and word embeddings, in conjunction with various classification algorithms, such as Naive Bayes, DTs, and SVM, with and without the use of SMOTE, is compared [12].

Evaluation of the performance of different SMOTE algorithm variants, such as Borderline-SMOTE, safe-level SMOTE, and SVM-SMOTE, with other oversampling techniques, such as random oversampling and SMOTE-adaptive is discussed. The effectiveness of these methods using a variety of classifiers, such as SVM, DTs, and RF is discussed [13]. Instead of randomly choosing neighbors from the full dataset like in the original SMOTE algorithm, the algorithm in [14] separates the minority class samples into several clusters and generates synthetic samples based on the nearest neighbors within the same cluster. The algorithm is called the cluster-SMOTE. A variety of classifiers, such as DTs, K-nearest neighbor (K-NN), and SVM, to compare the performance of the proposed algorithm with the original SMOTE algorithm and other oversampling techniques, such as ADASYN and Borderline-SMOTE (SVM) is used [14].

For the purpose of locating the minority class samples that require oversampling, Canopy and K-means clustering techniques are used. Based on how similar the data points are, the Canopy algorithm divides the dataset into a number of smaller subclusters, and the K-means method is then used in each subcluster to discover the minority class samples. Various classifiers, such as DTs, RFs, and SVM, to compare the performance of the proposed Canopy-K-means-SMOTE algorithm with that of the original SMOTE algorithm and alternative oversampling methods, such as ADASYN and Borderline-SMOTE (SVM) is discussed. According to the experimental findings, the suggested approach outperforms the original SMOTE algorithm and other oversampling methods in terms of a number of performance metrics, including accuracy, precision, recall, and F1-score [15]. Using the SMOTE process, a technique known as DeepSMOTE creates synthetic samples by first understanding the underlying distribution of the minority class using a deep learning model. It is discussed that using a deep learning model to generate synthetic samples can improve the quality of the synthetic samples by capturing the complex relationships between features and the class label [16]. Performance of various classifiers with and without the usage of SMOTE, including K-NN, DT (J48), RF, and SVM is depicted. The findings demonstrate that SMOTE greatly enhances the performance of all classifiers, especially for the minority class [17]. On a number of unbalanced datasets, SMOTE-based techniques are compared with a number of other cutting-edge oversampling and undersampling techniques. In order to learn more about how variables like dataset size, dimensionality, and class overlap affect classification performance, they also conduct data complexity analysis [18]. Using SMOTE and an extreme learning machine (ELM) classifier, a strategy for addressing unbalanced datasets is created. In order to oversample the minority class and balance the dataset, the suggested method first uses SMOTE. Then, an ELM classifier is trained using the oversampled dataset. Using a number of benchmark datasets, the performance evaluation of their suggested strategy in comparison to other cutting-edge approaches is carried out [19].

The imbalanced dataset problem is addressed using the genetic algorithm (GA) in SMOTE. Accordingly, the proposed method is superior to SMOTE without GA and other widely used techniques for handling imbalanced data in terms of F-measure, G-mean, and area under the curve (AUC) is discussed. The study also examines the suggested method's computational cost and demonstrates that it is comparable to existing methods [20]. The original SMOTE algorithm by generating synthetic instances for the minority class continuously, which allows it to adapt to changes in the data distribution over time is adapted. Performance assessment of C-performance SMOTEs on various benchmark datasets are compared to a number of other oversampling techniques for dynamic data streams. The findings demonstrate that, while being computationally efficient, C-SMOTE surpasses the other approaches in terms of accuracy, F1-score, and G-mean [21]. A class imbalance problem-containing NASA MDP dataset to train their prediction algorithms is considered. Several ML methods, such as logistic regression (LR), RF, and SVM, were evaluated for performance both with and without the use of SMOTE and grid search. The outcomes demonstrated that SMOTE and grid search considerably enhance the performance of the prediction models. The performance boost differs according to the ML technique employed, they also discovered [22]. By using the SMOTE algorithm, the issue of class imbalance in the intrusion detection dataset is solved. Ensemble approach that combines multiple classifiers to improve the overall performance of the system is used. The trials done reveal that their proposed methodology beats numerous existing methods in terms of accuracy, precision, recall, and F1 score [23]. To balance the unbalanced dataset of credit card transactions, the random over-sampling (ROS) is employed and SMOTE techniques. Using the original and the balanced dataset, the performance of four classifiers is assessed: LR, DTs, RF, and XGBoost. The results demonstrate the best accuracy, precision, recall, and F1-score performance for the SMOTE approach using the XGBoost classifier. The study finds that data sampling methods can greatly boost the effectiveness of systems that look for credit card fraud. However, detail about the drawbacks of strategy or the possible moral ramifications of employing data sampling techniques to detect credit card fraud is not discussed [24].

Variables associated with users' social behavior is extracted, including posting frequency, likes, and comments, using data from Twitter and Reddit. The social behavior of users has then been predicted using a variety of ensemble learning algorithms, including RF, AdaBoost, and gradient boosting [25]. Several combinations of characteristics and algorithms to discover the best-performing model is developed. They evaluated the performance of the models using metrics such as accuracy, precision, recall, and F1-score. The results reveal that ensemble learning algorithms can predict social behavior with great accuracy [25]. Financial fraud detection based on ensemble ML techniques is developed [26]. Three independent algorithms, namely XGBoost, RF, and Adaboost, is used to develop an ensemble model that integrates their individual predictions. SMOTE approach is employed to balance the classes in the study's unbalanced dataset. Using criteria like accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve, assessment of the performance of the suggested technique (AUC-ROC) is discussed. The findings demonstrate that the ensemble model outperforms the individual models and that the SMOTE technique enhances the models' performance. The suggested method may be helpful in identifying financial fraud and averting damages [26]. The study gives a comparative review of several ML algorithms for financial fraud detection, including DTs, RFs, gradient boosting, and SVMs. These methods are applied to a dataset of credit card transactions and evaluate their performance based on different metrics such as accuracy, precision, recall, and F1 score. To deal with the dataset's imbalance, they also test several sampling strategies, such as SMOTE and ADASYN. The results show that ensemble methods such as RFs and gradient boosting perform better than individual methods and that SMOTE and ADASYN can improve the performance of the models. The research provides useful insights into the selection of relevant machine-learning methods and sampling methodologies for financial fraud detection [27].

From the previous publication, it can be observed that much literature has proved that the classification is improved using the SMOTE and ensemble learning algorithms. This paper approaches the healthcare claim fraud detection classification problem using the SMOTE sampling method and the ensemble learning method. Different algorithms like LR, multilayer perceptron model, support vector classifier (SVC), and DT classifier are compared with the ensemble learning algorithm. The ensemble learning algorithm uses the metaclassifier as LR and base classifiers as SVC, multi-layer perceptron (MLP), LR, and DT algorithm. The stack ensemble learning method is applied and compared with the individual ML methods. This paper proposes an ensemble learning-based method for detecting healthcare claim fraud. With the ensemble learning algorithm proposed, the outcomes of different single learning methods, such as LR, DTs, MLP classifiers, and SVM are compared. A collection of health care claims that included data on the patient, the treating physician, and the services rendered is used. Outliers and missing values are removed during preprocessing of the dataset. To identify healthcare claim fraud, several single learning methods are used, such as LR, DTs, MLP classifiers, and SVM are compared with the ensemble learning algorithm in the imbalanced environment.

2. ENSEMBLE LEARNING-BASED HEALTHCARE FRAUD DETECTION USING SMOTE SAMPLING

Medicare is the healthcare welfare program from the US government. This program is exploited by fraudsters by means of fraudulent healthcare claims. The proposed implementation uses the dataset from Kaggle for healthcare claim fraud prediction in the present implementation [28]. The "Healthcare Fraud Detection" dataset is a set of fictitious medical claim data produced especially for the purpose of teaching ML models to recognize fraudulent claims. The dataset includes a range of data about healthcare claims, such as provider information: this includes information about the healthcare providers submitting the claims, such as their national provider identifier (NPI), address, and type of practice.

Information on the patients getting healthcare services, such as their gender, age, and location, is referred to as patient information. International classification of diseases (ICD) codes used to identify medical diagnoses connected to the healthcare claim are included in the diagnosis codes section. Current procedural terminology (CPT) codes, which are used to identify the medical procedures carried out as part of a healthcare claim, are included under the category of procedure codes. The payment amounts comprise both the whole amount billed for healthcare services as well as the sum received by the insurance company or other payers. The dataset was produced to aid in the development and testing of ML algorithms for detecting fraudulent medical claims. The data is constructed to resemble actual healthcare claims data, however, it does not actually contain any information about patients or providers due to its synthetic nature. As a result, scientists may test and improve their machine-learning models without endangering the privacy of real people. In general, the "Healthcare Fraud Detection" dataset is a useful tool for creating and evaluating ML models to identify fraudulent medical claims. Researchers can create models that can detect fraudulent claims based on a variety of different characteristics because of the dataset's breadth of information.

The complete data is split into training and testing data each having the inpatient, outpatient, beneficiary, and insurance provider data as separate CSV files. Insurance provider training data has the insurance provider's name with the information on whether there is fraud or not. Since the target attribute is whether the transaction is fraudulent or not the second column in the insurance provider test data will not have the fraudulent data column available since the ML algorithm has to predict the same. Inpatient data include beneficiary id, the amount reimbursed, claim start and end date, and the attended physician for the medical condition along with the provider data. Outpatient data include beneficiary id, the amount reimbursed, claim start and end date, diagnosis code, and the attended physician for the medical condition along with the provider data. Beneficiary data has personal information about the patient, beneficiary id annual deductible amount, reimbursement data, and medical condition of the patient. Outpatient and provider data are merged using the provider ID. Then this data frame is merged with the beneficiary data using the beneficiary id. Similarly, inpatient is merged with the provider data using provider id and then that data frame is merged with the beneficiary data with beneficiary id as the common column. Thus, two data sets with outpatient and inpatient are generated. These two are merged since they have similar column headings.

2.1. Data preprocessing and fraudulent transaction prediction

Since the data has many discrepancies before training it on the ML algorithm step by step data preprocessing method is applied to the merged dataset as given in the block diagram shown in Figure 1. The data required to develop the fraud detection algorithm is generated from different sources available and is discussed in detail in [28]. The data preprocessing starts with merging the data from different sources using the connecting attribute among the different data.

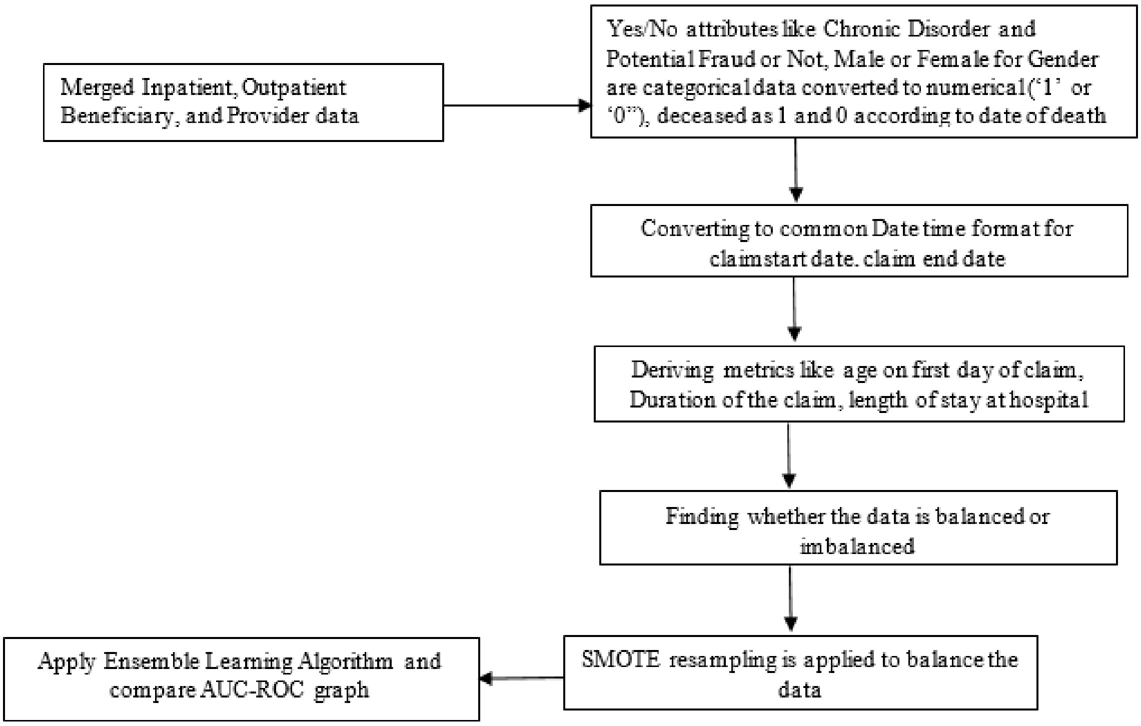

Merged Inpatient, Outpatient Beneficiary, and Provider data
Yes/No attributes like Chronic Disorder and Potential Fraud or Not, Male or Female for Gender are categorical data converted to numerical('1' or '0"), deceased as 1 and 0 according to date of death
Converting to common Date time format for claim start date. claim end date
Deriving metrics like age on first day of claim, Duration of the claim, length of stay at hospital
Finding whether the data is balanced or imbalanced
SMOTE resampling is applied to balance the data
Apply Ensemble Learning Algorithm and compare AUC-ROC graph

Figure 1. Flow diagram for ensemble learning implementation

After merging the data from all the sources including the inpatient, outpatient, beneficiary, and provider data the target attribute of potential fraud is checked for its count. After merging the data from different CSV files data preprocessing of the provided data is carried forward. The derived metrics including the age, and duration of the claim from the date of admission, and length of stay in the hospital are extracted. Data visualization to understand the data is carried out to get deeper insights into the fraudulent and nonfraudulent data.

3. RESULTS AND DISCUSSION

Healthcare claim fraud detection is implemented using Python with sklearn, SMOTE, and pandas toolboxes. Ensemble learning code is developed for healthcare claim fraud detection using LR as the meta classifier and MLP, SVM, and DT algorithm as the base classifier. The bar graph drawn depicting the class count for both fraudulent and non-fraudulent data is shown in Figure 2. It can be observed that the fraudulent data based on the renal disease data seems to be imbalanced as shown in Figure 3.

For further scrutiny of the data, some parameters are checked for in the idea about the fraudulent transactions. Chronic kidney disease is at large responsible for renal disease. An observation in the dataset reveals a few important aspects that clearly indicates a higher possibility of fraudulent transactions. The bar graph in Figure 3 indicates the count of people who are affected with renal disease, chronic kidney disease, and renal disease without chronic kidney disease. The count of renal disease without chronic kidney disease is a doubtful area to be scrutinized. A lot of fraudulent transactions would have occurred in this area. Figure 3 depicts three important counts that define the chronic and renal disease count in the dataset and the patient's claim that renal disease has occurred without them having chronic kidney ailments previously.

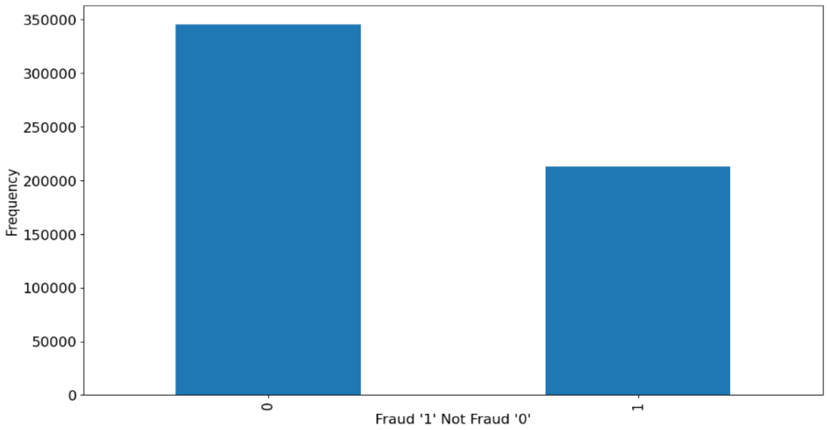


Figure 2. Potential fraud ‘1’ for fraudulent transaction and ‘0’ for not a fraudulent transaction

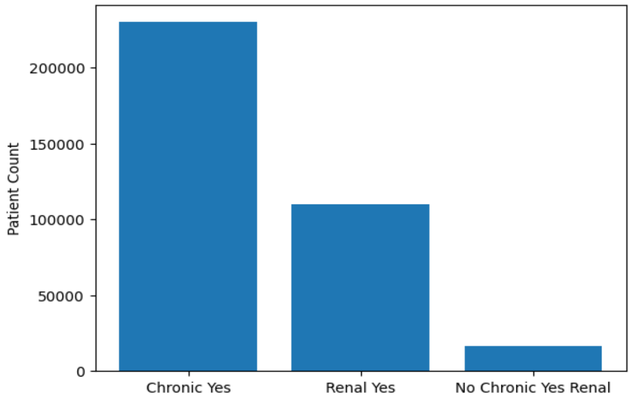
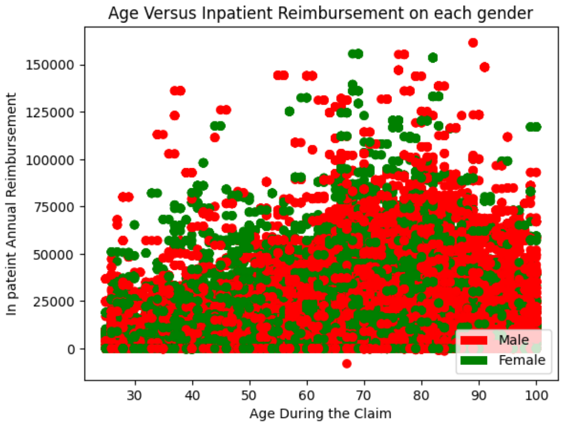


Figure 3. Disease versus count (kidney disease)

These instances where patients have renal diseases without having chronic kidney disease can be fraudulent transactions due to suspicion. Inpatient and outpatient reimbursement in the different age groups has to be scrutinized whether a particular age group is involved in a fraudulent transaction. This plot is plotted highlights each gender as shown in Figure 4. On observing the scatter plot, it is evident that age and gender do not affect the in-patient annual reimbursement and thus indicating that fraudulent transactions are not dependent on either the age or gender of the patient. Both the inpatient and outpatient reimbursement amount data are scattered equally in the age graph. Thus, these attributes of age and reimbursement amount do not come into suspicion as depicted in Figure 5.



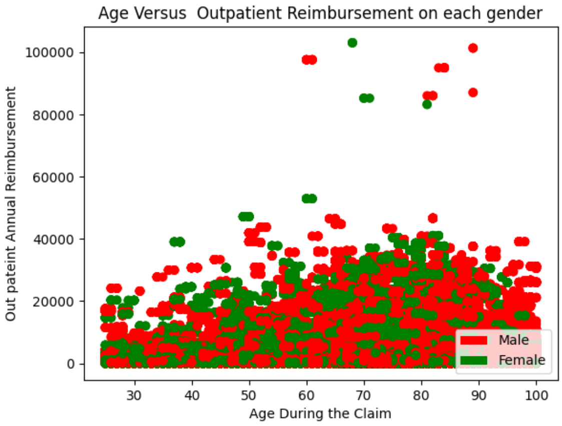


Figure 4. In patient reimbursement data

Figure 5. Outpatient reimbursement data

The classification problem is formulated using individual and ensemble learning paradigms. Data is checked for data imbalance and data is resampled. Before resampling, many attributes that are not relevant for classification are dropped from the dataset. Attributes that are removed include beneficiary ID, claim ID, date of death, claim diagnostic code, and details about the physician. After removing the unimportant features, the important features from the dataset are extracted using the recursive feature elimination method. The parameters used for the different ML algorithms used in the ensemble learning method are given in Table 1.

Table 1. Ensemble learning parameters

ML model	Parameters
MLP classifier	activation="relu", alpha=0.1, hidden_layer_sizes=(20,20,20), learning_rate="constant", max_iter=3000, random_state=1000
Decision tree Classifier	max_depth=25, max_features="auto", min_samples_leaf=0.005, min_samples_split=0.005, random_state=2000
SVC	C=1, degree=1, gamma=.1, kernel="poly", probability=true
Logistic regression	random_state=42

Ensemble learning is a ML technique where multiple base classifiers are combined to create a stronger predictive model. In this case, the base classifiers are SVC, MLP, and DT algorithm. The first step in the ensemble learning algorithm is to train each of the base classifiers independently on the training data. Each base classifier will produce its own set of predictions for the test data. These predictions will be combined to make a final prediction using a meta-classifier. The meta classifier in this case is LR. It takes the predictions from each of the base classifiers as input and uses them to make a final prediction. The meta-classifier learns from the predictions made by the base classifiers and combines them in a way that produces the most accurate prediction possible.

The final ensemble model is created by combining the predictions from the base classifiers with the predictions made by the meta-classifier. This combination is done in a way that maximizes the accuracy of the predictions. The advantage of using an ensemble learning algorithm is that it can improve the accuracy of the model by combining the strengths of multiple base classifiers. Each base classifier has its own strengths and weaknesses, and by combining them, the weaknesses of one classifier can be offset by the strengths of another. Overall, the ensemble learning algorithm using SVC, MLP, and DT algorithm as base classifiers and LR as a meta-classifier is a powerful ML technique that can produce highly accurate predictions.

The AUC-ROC curve obtained from the individual and ensemble learning (stack distribution) algorithm is given in Figure 6. The performance of the AUC-ROC curve is best obtained from the stack distribution algorithm as compared to the other individual ML algorithms as shown in Figure 6. Among the individual algorithms, LR performed well. The ensemble learning algorithm outperforms all the individual algorithms. Since the sampling method developed is SMOTE algorithm the accuracy of the implementation is found to be satisfactory. Since there are different SMOTE algorithms available the variation of SMOTE implementation on upsampling can be exploited for enhancing the accuracy and AUC-ROC curve for the upsampled input attributes.

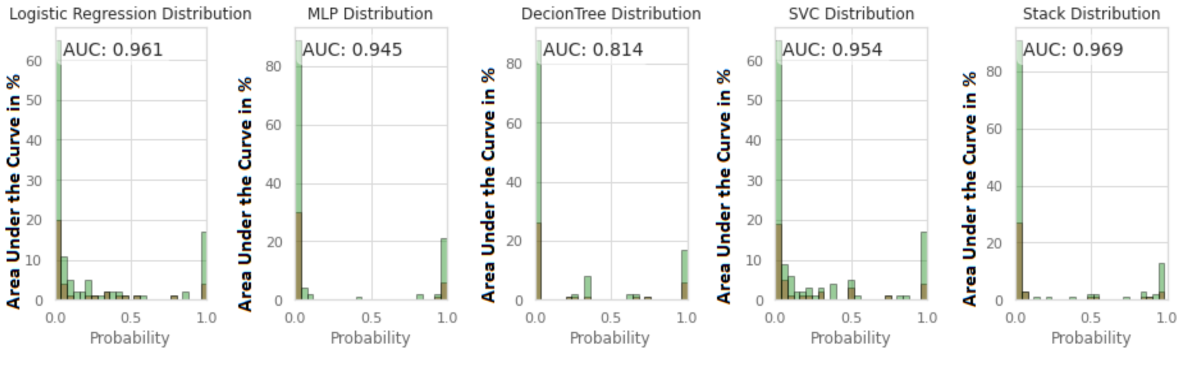


Figure 6. AUC-ROC curve for individual and stack ensemble learning algorithm

4. CONCLUSION

In this paper, an ensemble learning-based approach for the detection of healthcare claim fraud is developed. We explored the performance of various ensemble methods and compared them with various single-learning algorithms. Our results demonstrate that the ensemble learning approach significantly outperforms single learning algorithms for the detection of healthcare claim fraud. Our approach achieved an accuracy of 97%, indicating the potential for efficient fraud detection in the healthcare domain. Future work includes the exploration of other ensemble learning algorithms and the extension of the approach to other domains. Our experimental results show that the ensemble learning approach significantly outperforms single learning algorithms for the detection of healthcare claim fraud.

REFERENCES

[1] J. Lu, K. Lin, R. Chen, M. Lin, X. Chen, and P. Lu, "Health insurance fraud detection by using an attributed heterogeneous information network with a hierarchical attention mechanism," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, p. 62, Apr. 2023, doi: 10.1186/s12911-023-02152-0.

[2] I. Matloob, S. A. Khan, R. Rukaiya, M. A. K. Khattak, and A. Munir, "A sequence mining-based novel architecture for detecting fraudulent transactions in healthcare systems," *IEEE Access*, vol. 10, pp. 48447–48463, 2022, doi: 10.1109/ACCESS.2022.3170888.

[3] S. Zhou, J. He, H. Yang, D. Chen, and R. Zhang, "Big data-driven abnormal behavior detection in healthcare based on association rules," *IEEE Access*, vol. 8, pp. 129002–129011, 2020, doi: 10.1109/ACCESS.2020.3009006.

[4] Y. Gao, C. Sun, R. Li, Q. Li, L. Cui, and B. Gong, "An efficient fraud identification method combining manifold learning and outliers detection in mobile healthcare services," *IEEE Access*, vol. 6, pp. 60059–60068, 2018, doi: 10.1109/ACCESS.2018.2875516.

[5] M. E. Haque and M. E. Tozal, "Identifying health insurance claim frauds using mixture of clinical concepts," *IEEE Transactions on Services Computing*, vol. 15, no. 4, pp. 2356–2367, Jul. 2022, doi: 10.1109/TSC.2021.3051165.

[6] F. Kamalov, A. Elnagar, and H. H. Leung, "Ensemble learning with resampling for imbalanced data," in *International Conference on Intelligent Computing*, 2021, pp. 564–578, doi: 10.1007/978-3-030-84529-2_48.

[7] Y. Yang, P. Xiao, Y. Cheng, W. Liu, and Z. Huang, "Ensemble strategy for hard classifying samples in class-imbalanced data set," in *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Jan. 2018, pp. 170–175, doi: 10.1109/BigComp.2018.00033.

[8] W. D. Alnatara and M. L. Khodra, "Imbalanced data handling in multi-label aspect categorization using oversampling and ensemble learning," in *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Oct. 2020, pp. 165–170, doi: 10.1109/ICACSIS51025.2020.9263087.

[9] W. Wan, Q. Xu, H. Chen, and Q. Chen, "Using deep learning neural networks and stacking ensemble learning to predict CSI 300 index," in *2022 9th International Conference on Digital Home (ICDH)*, Oct. 2022, pp. 81–86, doi: 10.1109/ICDH57206.2022.00020.

[10] Y. Cheng, Y. Li, H. Wu, F. Li, Y. Li, and L. He, "Soil moisture retrieval using stacked generalization: an ensemble machine learning method," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, Jul. 2021, pp. 6984–6987, doi: 10.1109/IGARSS47720.2021.9554608.

[11] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, "SMOTE for handling imbalanced data problem : a review," in *2021 Sixth International Conference on Informatics and Computing (ICIC)*, Nov. 2021, pp. 1–8, doi: 10.1109/ICIC54025.2021.9632912.

[12] W. Satriaji and R. Kusumaningrum, "Effect of synthetic minority oversampling technique (SMOTE), feature representation, and classification algorithm on imbalanced sentiment analysis," in *2018 2nd International Conference on Informatics and Computational Sciences (ICICoS)*, Oct. 2018, pp. 1–5, doi: 10.1109/ICICOS.2018.8621648.

[13] D. Bajer, B. Zonc, M. Dudjak, and G. Martinovic, "Performance analysis of SMOTE-based oversampling techniques when dealing with data imbalance," in *2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*, Jun. 2019, pp. 265–271, doi: 10.1109/IWSSIP.2019.8787306.

[14] Z. Hengyu, "Improved SMOTE algorithm for imbalanced dataset," in *2020 Chinese Automation Congress (CAC)*, Nov. 2020, pp. 693–697, doi: 10.1109/CAC51589.2020.9326603.

[15] C. Guo, Y. Ma, Z. Xu, M. Cao, and Q. Yao, "An improved oversampling method for imbalanced data–SMOTE based on Canopy and K-means," in *2019 Chinese Automation Congress (CAC)*, Nov. 2019, pp. 1467–1469, doi: 10.1109/CAC48633.2019.8997367.

[16] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: fusing deep learning and SMOTE for imbalanced data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 9, pp. 6390–6404, Sep. 2023, doi: 10.1109/TNNLS.2021.3136503.

[17] A. A. Alfrhan, R. H. Alhusain, and R. U. Khan, "SMOTE: class imbalance problem in intrusion detection system," in *2020 International Conference on Computing and Information Technology (ICCIT-1441)*, Sep. 2020, pp. 1–5, doi: 10.1109/ICCIT-144147971.2020.9213728.

[18] N. A. Azhar, M. S. M. Pozi, A. M. Din, and A. Jatowt, "An investigation of SMOTE based methods for imbalanced datasets with data complexity analysis," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2022, doi: 10.1109/TKDE.2022.3179381.

[19] R. Rustogi and A. Prasad, "Swift imbalance data classification using SMOTE and extreme learning machine," in *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, Feb. 2019, pp. 1–6, doi: 10.1109/ICCIDS.2019.8862112.

[20] T. E. Tallo and A. Musdholifah, "The implementation of genetic algorithm in smote (synthetic minority oversampling technique) for handling imbalanced dataset problem," in *2018 4th International Conference on Science and Technology (ICST)*, Aug. 2018, pp. 1–4, doi: 10.1109/ICSTC.2018.8528591.

[21] A. Bernardo, H. M. Gomes, J. Montiel, B. Pfahringer, A. Bifet, and E. D. Valle, "C-SMOTE: continuous synthetic minority oversampling for evolving data streams," in *2020 IEEE International Conference on Big Data (Big Data)*, Dec. 2020, pp. 483–492, doi: 10.1109/BigData50022.2020.9377768.

[22] E. Sara, C. Laila, and I. Ali, "The impact of SMOTE and grid search on maintainability prediction models," in *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, Nov. 2019, pp. 1–8, doi: 10.1109/AICCSA47632.2019.9035342.

[23] F. Li, W. Ma, H. Li, and J. Li, “Improving intrusion detection system using ensemble methods and over-sampling technique,” in *2022 4th International Academic Exchange Conference on Science and Technology Innovation (IAECST)*, Dec. 2022, pp. 1200–1205, doi: 10.1109/IAECST57965.2022.10062178.

[24] T. J. Berkmans and S. Karthick, “Credit card fraud detection with data sampling,” in *2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)*, Dec. 2022, pp. 1–6, doi: 10.1109/ICPECTS56089.2022.10046729.

[25] M. Raihan, N. Alvi, and A. K. Bairagi, “Social behavior prediction using ensemble learning algorithms,” in *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Oct. 2022, pp. 1–5, doi: 10.1109/ICCCNT54827.2022.9984251.

[26] J. Wang and C. Yang, “Financial fraud detection based on ensemble machine learning,” in *2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, Sep. 2022, pp. 1–6, doi: 10.1109/DASC/PiCom/CBDCom/Cy55231.2022.9928001.

[27] A. Menshchikov, V. Perfilev, D. Roenko, M. Zykin, and M. Fedosenko, “Comparative analysis of machine learning methods application for financial fraud detection,” in *2022 32nd Conference of Open Innovations Association (FRUCT)*, Nov. 2022, pp. 178–186, doi: 10.23919/FRUCT56874.2022.9953872.

[28] R. A. Gupta, “Healthcare provider fraud detection analysis,” 2021. https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis.

BIOGRAPHIES OF AUTHORS

Mrs. Shweta S. Kaddi is presently working as an Assistant Professor in the Department of Computer Science and Engineering at J.S.S. Academy of technical education, Bangalore, Karnataka, India. She received her M.Tech. from Visvesaraya Technical University, Belgavi in the year 2010. Her research interests are clinical data mining, machine learning, and artificial intelligence. She can be contacted at email: shwetakaddi@gmail.com.

Dr. Malini M. Patil has 26 years of academic and 15 years of research experience. Currently, she is working as a Professor and Head of the Department at R V Institute of Technology and Management, Bengaluru, Karnataka, India. Her research interests are data mining, machine learning, and big data analytics. She has published more than 60 research articles in reputed international journals and conferences in India and Abroad. Presently she is guiding three research Scholars, Three of her scholars were awarded with Ph.D. degrees. She has authored two book chapters. She is a member of professional societies, a senior member of IEEE, a member of Women in Engineering (WIE), LMCSI, LMIEI, LMISTE, a Member of ACM, and actively involved in all professional society activities. She has delivered many webinars /seminars as an invited speaker across India. She has visited Hongkong, Dubai, Srilanka, and Malaysia to attend and present her research papers at international conferences. She received the best paper presentation award at many conferences. She has been a technical and advisory committee member of many international conferences She has organized several FDP, conferences, workshops, and seminars and chaired many sessions at international conferences. She obtained her UG degree from Karnataka University, Dharwad, her PG degree from Shivaji University, Kolhapur, and her Ph.D. from Bharathiar University, Tamilnadu. She can be contacted at email: drmalinimpatil@gmail.com.