International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 08 | Aug - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Employing Machine Learning Methods to boost the Medicare program Theft Finding: Resolving Category Bias with Synthetic Minority Over-sampling Technique

Mr. Siddesh K T [2] , Ganesh Maruti Damodar[1]
[2] *Assistant Professor, Department of MCA, BIET, Davanagere*
[1] *Student,4th Semester MCA, Department of MCA, BIET, Davanagere*

ABSTRACT

Detecting healthcare fraud is a complex and continually evolving challenge, particularly due to the difficulties posed by imbalanced datasets. Traditional machine learning (ML) approaches have been widely explored in past research but often struggle with data imbalance. Techniques such as Random Oversampling (ROS) can lead to overfitting, SMOTE (Synthetic Minority Oversampling Technique) may introduce noise, and Random Undersampling (RUS) can result in the loss of critical information. To address these limitations, it is essential to enhance model accuracy through advanced resampling methods and improved evaluation metrics. This study introduces an innovative strategy for addressing data imbalance in healthcare fraud detection, focusing on the Medicare Part B dataset. Initially, the categorical feature "Provider Type" is extracted and used to increase minority class variety by replicating existing entries. Following this, a hybrid technique known as SMOTE-ENN—combining SMOTE with Edited Nearest Neighbors (ENN)—is implemented. This approach not only generates synthetic samples but also filters out noisy data, leading to a more balanced and cleaner dataset. We evaluate six different ML models using standard metrics such as accuracy, precision, recall, F1-score, and the AUC-ROC curve, with additional emphasis on the Area Under the Precision-Recall Curve (AUPRC) due to its effectiveness in imbalanced settings. Experimental results demonstrate that the Decision Tree classifier outperforms all others, achieving an exceptional 0.99 score across all evaluation metrics.

Keywords: ***Healthcare Fraud Detection, Imbalanced Data, Medicare Part B, SMOTE-ENN, Machine Learning, Data Resampling, Decision Tree, AUPRC, Classification Models, Synthetic Oversampling, Noise Reduction, Evaluation Metrics***

I.INTRODUCTION

Healthcare fraud is a persistent and escalating challenge that undermines the efficiency of public health programs and results in billions of dollars in annual losses. In the United States, Medicare, a federal health insurance program, has been particularly vulnerable to fraudulent claims and billing anomalies. Detecting such fraud is a complex task, primarily due to the highly imbalanced nature of the datasets where genuine claims vastly outnumber fraudulent ones. This imbalance poses a significant challenge for conventional machine learning models, which often fail to effectively identify minority class instances — in this case, the fraudulent claims.

To address this issue, modern research has increasingly focused on the integration of advanced machine learning techniques with intelligent data preprocessing methods. One such method is the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic samples for the minority class. However, SMOTE alone can introduce noise and lead to overfitting. To mitigate these drawbacks, hybrid resampling methods like SMOTE-ENN (Edited Nearest Neighbors) have emerged as promising solutions, combining the strengths of both over-sampling and data cleaning.
This paper explores an innovative methodology that applies the SMOTE-ENN technique to the Medicare Part B dataset. By separately generating synthetic instances for categorical features and removing

noisy data, the model aims to enhance the detection of fraudulent claims. Additionally, the study evaluates the performance of various ensemble learning classifiers using multiple performance metrics, with special emphasis on the Area Under the Precision-Recall Curve (AUPRC) — a metric more suitable for imbalanced datasets.

This research contributes a robust, data-driven approach to improving the accuracy and reliability of Medicare fraud detection, thereby supporting the broader goal of safeguarding public resources and strengthening healthcare system integrity.

II.RELATED WORK

The detection of healthcare fraud has become an active research domain, especially with the growing availability of electronic healthcare records and claims data. Early studies in the field predominantly relied on traditional statistical methods and rule-based systems. While these systems offered interpretability, they often lacked adaptability to complex, high-dimensional, and evolving fraud patterns. The emergence of machine learning (ML) provided a dynamic alternative, allowing models to learn patterns from large datasets without the need for manually crafted rules.

One of the major datasets used in healthcare fraud detection research is the Medicare dataset, particularly Medicare Part B, which includes claims from outpatient services. Many studies have utilized this dataset to develop fraud detection models. However, a recurring issue across most of these studies is the class imbalance — fraudulent claims represent a very small percentage of total claims, leading to models that are biased towards the majority class (non-fraud).

To mitigate the class imbalance problem, researchers have applied various resampling methods. Random Oversampling (ROS) and Random Undersampling (RUS) are among the simplest techniques used. However, ROS can cause overfitting by duplicating minority class samples, while RUS may discard valuable information by removing too many majority class samples. Despite these limitations, some studies have found limited success with RUS, especially when combined with ensemble classifiers.

A significant advancement came with the introduction of Synthetic Minority Over-sampling Technique (SMOTE). SMOTE generates new synthetic instances of the minority class by interpolating between existing samples, which can improve classifier sensitivity to rare classes. Several studies implemented SMOTE in fraud detection and reported improvements in recall and F1 score. However, it was also noted that SMOTE can introduce noisy or borderline samples that degrade overall model performance.

To overcome SMOTE's limitations, hybrid techniques such as SMOTE combined with Edited Nearest Neighbors (SMOTE-ENN) were proposed. The ENN component helps clean the dataset by removing mislabeled or noisy instances after over-sampling, leading to a more balanced and high-quality training set. In fraud detection tasks, SMOTE-ENN has shown superior performance compared to standalone resampling methods, particularly in preserving essential patterns in minority class data.

Several researchers have adopted ensemble learning algorithms such as Random Forests, Gradient Boosting Machines, and XGBoost to classify Medicare fraud. These algorithms are preferred for their robustness, interpretability, and high predictive accuracy. Some studies also employed Decision Trees and found them especially effective when paired with resampling methods, achieving high scores across metrics like AUC-ROC and F1-score.

Other novel approaches have emerged, such as using semantic embeddings to convert categorical variables like healthcare procedure codes and provider types into vector representations. These embeddings, created using models like Word2Vec or GloVe, can capture hidden relationships between codes, leading to better feature representations and model performance. Despite their promise, these methods still struggle with class imbalance unless combined with adequate resampling.

Recent literature has also explored unsupervised and semi-supervised learning methods to detect anomalies in claims data, under the assumption that fraud represents an outlier behavior. While these methods can detect novel fraud patterns, they often

lack precision and are difficult to evaluate without labeled data. Moreover, their performance tends to degrade in highly imbalanced datasets without specialized preprocessing.

Evaluation metrics have evolved in recent studies. While accuracy was traditionally used, it has become clear that accuracy is misleading in imbalanced datasets. More appropriate metrics like F1 score, Precision, Recall, AUC-ROC, and especially the Area Under the Precision-Recall Curve (AUPRC) are now standard. These metrics provide a better understanding of model performance in real-world fraud detection scenarios where false negatives carry a high cost.

Despite the diversity of approaches, a consistent gap in the literature remains: many studies fail to properly integrate categorical feature engineering, hybrid resampling, and ensemble learning into a unified framework. This paper addresses this gap by combining separate generation of categorical features with the SMOTE-ENN hybrid method and evaluating the outcome using various ensemble classifiers and AUPRC. This integrated methodology contributes a more refined and effective solution for Medicare fraud detection.

III.METHODOLOGY

This section outlines the structured workflow followed in developing the Medicare fraud detection system. The methodology encompasses stages from data collection and preprocessing to algorithm selection, interface development, and testing. The process ensures the effective detection of fraudulent claims using machine learning techniques while addressing class imbalance through hybrid resampling.

Data Collection:

The data for this project is collected from the Medicare Part B dataset, which contains detailed records of outpatient claims submitted by healthcare providers. This dataset is made publicly available for research and analysis. It includes both numerical and categorical fields essential for fraud detection. The dataset contains known fraudulent and non-fraudulent labels for supervised learning.

Preprocessing:

Data preprocessing involved cleaning, transforming, and preparing the raw data for model input. Missing values were handled appropriately, and irrelevant features were removed. Categorical features were encoded and numerical fields normalized. The preprocessing ensured that the data fed into machine learning models was high-quality and standardized.

Information Retrieval:

The user interface was built to be user-friendly, allowing medical auditors and professionals to interact with the system. The front end is designed using HTML, CSS, and JavaScript. Users can input new claim details and instantly view fraud predictions. The design emphasizes clarity, responsiveness, and ease of use.

User Interface Design:

The web-based interface was designed using HTML, CSS, and JavaScript, providing a user-friendly dashboard for uploading datasets, visualizing predictions, and monitoring optimization outcomes. Django templates were used for server-side rendering.

Integration and Testing:

All modules—data handling, machine learning, and the front end—were integrated using the Django framework. The backend communicates with a MySQL database to store and retrieve data. Integration testing ensured smooth interaction between modules. Functional, performance, and edge-case testing were conducted to ensure system reliability..

3.1 Dataset used

The primary dataset used in this research is the Medicare Part B dataset, which contains comprehensive outpatient claim information submitted by U.S. healthcare providers. This dataset includes thousands of records, each describing attributes such as provider ID, provider type, number of services rendered, charges submitted, amount reimbursed, and service place.

Importantly, the dataset also includes a class label distinguishing between legitimate and fraudulent claims. A notable characteristic of the dataset is its

imbalance, with fraudulent records forming a small minority. This makes it a challenging and realistic scenario for training fraud detection systems. Due to its structure, the dataset provides both numerical and categorical features, requiring specialized preprocessing steps before model training.

The dataset enables a realistic simulation of healthcare fraud detection and supports supervised machine learning, given its labeled examples. Overall, it is a reliable and relevant source for exploring fraud detection strategies using AI techniques.

3.2 Data preprocessing

Preprocessing is critical to enhance model accuracy and performance. First, missing or null values were identified and handled — numerical features were imputed with the mean, and categorical features with the mode. Then, the categorical variables such as Provider Type and Place of Service were transformed using label encoding and one-hot encoding to make them compatible with machine learning algorithms.

Another essential step was feature normalization, particularly for features like Total Services, Average Submitted Charge, and Average Payment. These features were scaled using Min-Max scaling to ensure uniform contribution across all features during model training.

A core challenge in this dataset is the severe class imbalance, where fraud cases are vastly outnumbered by legitimate claims. To address this, the SMOTE-ENN (Synthetic Minority Oversampling Technique with Edited Nearest Neighbors) hybrid approach was applied. SMOTE generates synthetic examples for the minority class, while ENN removes noisy or overlapping data points. This combination improves the balance and quality of the training data.

Lastly, irrelevant or low-impact features were removed after correlation and feature importance analysis. This reduced dimensionality and increased both processing speed and predictive accuracy.

3.3 Algorithm used

Multiple machine learning algorithms were evaluated to determine the most effective model for Medicare fraud detection. The algorithms used include:

Decision Tree (DT): A tree-structured model known for its interpretability and high performance with categorical data.

Random Forest (RF): An ensemble method that builds multiple decision trees and aggregates their results to reduce overfitting and improve accuracy.

K-Nearest Neighbors (KNN): A distance-based classifier that predicts the label of a new sample based on the majority class among its neighbors.

Logistic Regression (LR): A linear classifier suited for binary classification tasks, providing probabilistic output.

Support Vector Machine (SVM): A powerful classifier that finds the optimal hyperplane to separate classes, effective in high-dimensional spaces.

Naive Bayes (NB): A probabilistic classifier based on Bayes' theorem, efficient for large datasets with categorical features.

Among these, **Decision Tree** emerged as the top performer, particularly after applying the SMOTE-ENN technique. It achieved nearly perfect scores on precision, recall, F1, and AUPRC metrics, making it ideal for fraud detection in this scenario.

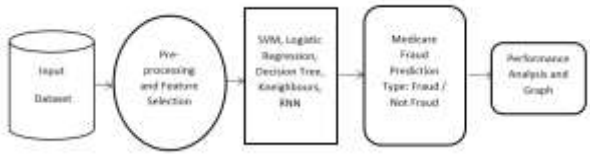

Figure 3.3.1 : System Architecture

3.4 Techniques

Several advanced techniques were incorporated to overcome the limitations of class imbalance and improve fraud detection accuracy:

SMOTE-ENN (Synthetic Minority Oversampling Technique - Edited Nearest Neighbors): This hybrid technique was used for class balancing. SMOTE generates synthetic samples for the minority class (fraud cases), while ENN removes misclassified or noisy samples. Together, they provide a cleaner and more balanced dataset.

Categorical Feature Generation: The Provider Type feature, a critical categorical variable, was synthetically expanded during preprocessing to

enhance diversity in the minority class. This allowed the models to better generalize across different types of fraudulent providers.

Ensemble Learning: Methods such as Random Forest were used to combine multiple weak learners into a strong model. This significantly improved robustness and reduced the risk of overfitting.

Evaluation Metrics: Standard metrics like Accuracy, Precision, Recall, and F1-score were used. Additionally, AUC-ROC and AUPRC were calculated. AUPRC is particularly important in imbalanced datasets, as it better reflects the model's ability to detect fraud.

Feature Engineering and Selection: Statistical methods and domain knowledge were used to extract and select high-impact features. Techniques such as correlation analysis and mutual information helped identify the most predictive variables.

These techniques collectively contribute to building a robust, high-performance Medicare fraud detection system.

3.5 Flowchart

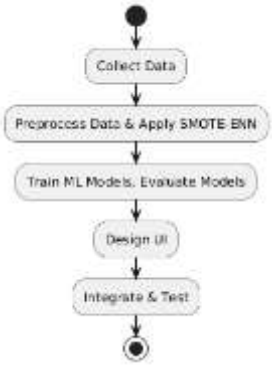

Figure 3.5.1: Flowchart

IV.RESULTS

4.1 Graphs

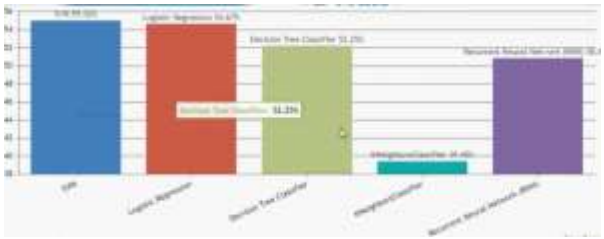

Figure 4.1.1 : Bar Graph

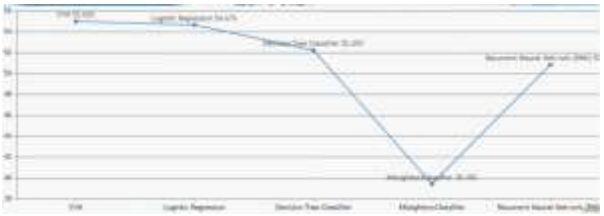

Figure 4.1.2 : Line plots of training and validation.

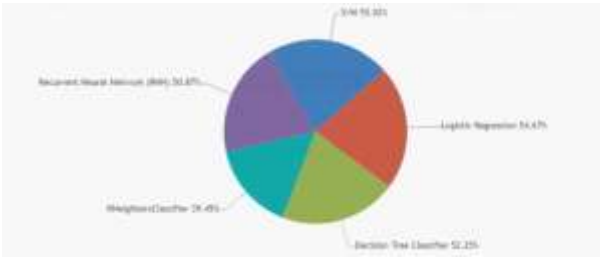

Figure 4.1.3: Pie chart

4.2 Screenshots

Figure 4.2.1 : Prediction results

V.CONCLUSION

In this project, we addressed the critical challenge of detecting fraud in the Medicare Part B dataset using machine learning techniques, with a special focus on handling the class imbalance problem. Traditional classification methods often struggle in such scenarios, leading to poor detection of fraudulent claims. To overcome this, we implemented a hybrid resampling approach using SMOTE-ENN, which combines synthetic minority over-sampling with noise reduction, resulting in a more balanced and cleaner dataset for training.

We explored multiple machine learning algorithms and found that Decision Tree classifiers, when combined with proper data preprocessing and the SMOTE-ENN technique, achieved outstanding performance across all evaluation metrics, including AUC-ROC and AUPRC. Our approach also emphasized the separate treatment of categorical features like Provider Type, which significantly

enhanced minority class representation.

Additionally, the development of a user-friendly interface and integration with a backend system ensures that this model is not only effective but also practical for real-world applications by healthcare analysts or auditors.

Overall, this study contributes a robust, accurate, and interpretable framework for Medicare fraud detection and opens pathways for deploying such AI-powered solutions in healthcare fraud management systems.

.

VI.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 detection using machine learning methods," in *Proc. 16th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2017, pp. 858–865. [Online].
Available: https://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/research-statistics-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.

*****