National College of Ireland

Advanced Resampling Techniques and Ensemble Methods for Improved Detection of Imbalanced Healthcare Fraud Cases

MSc Research Project
MSc in Data Analytics

Snovy Lopes
Student ID: 23242221

School of Computing
National College of Ireland

Supervisor: Cristina Hava Muntean

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Snovy Lopes
Student ID:	23242221
Programme:	MSc in Data Analytics **Year:** 2024-25
Module:	MSc Research Project
Supervisor:	Cristina Hava Muntean
Submission Due Date:	12 December 2024
Project Title:	Advanced Resampling Techniques and Ensemble Methods for Improved Detection of Imbalanced Medicare Fraud Cases
Word Count:	**Page Count: 20 pages**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Snovy Lopes

Date: 12-12-2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Advanced Resampling Techniques and Ensemble Methods for Improved Detection of Imbalanced Healthcare Fraud Cases

Snovy Lopes
23242221

Abstract

Healthcare fraud detection is crucial due to the large amount of financial losses that occur due to it varying from 3% to 10% of the total income or 19 to 65 billion USD a year in the U. S Medicare system. Traditional rule-based fraud detection systems has proven to be insufficient in the current dynamic environment of fraud. However, machine learning (ML) provides potential solutions for designing more effective detection algorithms which is especially noticeable in work with large datasets. But the challenge of imbalanced datasets persists. This study examines the SMOTE-ENN hybrid resampling method with other superior methods and then compares the accuracy of the ensemble methods with the single models. In this study, Random Forest with SMOTE-ENN is compared with the other methods and evaluated using metrics like precision, recall, F1 score, AUC-PR curve, AUC-ROC curve, confusion matrix, etc. The resulting conclusion is SMOTE-ENN with Random Forest performing better compared to the other combinations.

Keywords—**SMOTE-ENN, Random forest, Machine learning, Class imbalance, Healthcare fraud, Ensemble methods, Resampling techniques.**

1 Introduction

Fraud in the Medicare sector is a widespread and growing problem that affects both the financial and overall credibility of the healthcare provision chain and it is increasing in proportion every year (Bauder, 2017). This type of fraud involves deliberate deception to secure an unfair or unlawful gain from healthcare programs, whether it be government-funded or private insurance schemes Fraudulent activities in healthcare include, but are not limited to, billing for services not provided, misrepresenting the cost of services, and providing unnecessary medical services or treatments. Specifically, the U.S. Medicare program is one of the largest victims of healthcare fraud, with damages ranging from 19 to 65 billion dollars annually (Bounab, 2024). It is not only a concern for the facilities and staff directly exposed to it but also a problem that the entire healthcare system suffers from, despite the fact that resources and efforts are being used up in managing it instead of investing them in tangible patient care (Bauder, 2018).

In response to these growing challenges, healthcare organizations have turned to advanced technological solutions, such as Machine Learning (ML) which is a subfield of Artificial Intelligence (AI) to enhance fraud detection mechanisms (J. T. Hancock, 2023). The traditional anti-fraud systems have relied heavily on rule-based methods which are effective to some degree, but they are limited by their inability to adapt to the ever-evolving nature of fraud schemes. Fraudulent actors continuously develop new tactics, rendering static rule-based systems insufficient for detecting emerging patterns of fraud [Cheah et al., 2023]. This limitation has led to the increased adoption of ML techniques like Random Forest, XGBoost and Support Vector Machine (SVM), etc., which offer more flexible and robust solutions capable of identifying complex patterns and learning from large datasets (Bauder, 2017) (Hancock, 2023).

The Centers for Medicare and Medicaid Services (CMS) has made Medicare data available to the public to help fight healthcare fraud. For this case study, we use the 2022 Medicare Part D Prescribers - by Provider and Drug which will be henceforth called as Medicare Part D (Anon., 2022). This dataset provides information on prescription drugs provided to Medicare beneficiaries by physicians and other health care providers. The Medicare Part D dataset does not include a fraud label, but one can cross-reference it with the Office of Inspector General's (OIG) List of Excluded Individuals/Entities (LEIE) database (Anon., n.d.). This list includes healthcare providers who have been banned from participating in Medicare due to misconduct for a certain period of time. However, it's important to note that even though providers may be on this list, 38% of those with fraud convictions still practice medicine, and 21% were not suspended despite their convictions, meaning the list does not capture all fraud cases (Maas, 2013). Despite these limitations, using large datasets like these, combined with machine learning to detect fraud, could help recover significant costs for Medicare.

Surveys of existing literature reveal that, though ML techniques have promising accuracy, the problem of imbalanced data where the number of fraudulent cases is considerably smaller than non-fraudulent cases persist as a major issue (Bauder, 2018) (Bauder, 2018) (Hancock, 2022). Most of the real-world Medicare datasets suffer from severe class imbalance issues. The class imbalance in Medicare datasets makes the model more biased on predicting non-fraudulent claims, causing it to miss many fraudulent ones and perform poorly in detecting fraud (Bauder, 2018). This research addresses the problem of class imbalance by using hybrid resampling techniques like Synthetic Minority Oversampling Technique with Edited Nearest Neighbors (SMOTE-ENN) and SMOTE-Tomek. The **research question** of this study is as followed– How does the performance of the SMOTE-ENN hybrid resampling technique compared to SMOTE-Tomek in detecting healthcare fraud using a Random Forest classifier, evaluated based on accuracy, precision, recall, F1-score and AUC-ROC metrics?

The research objectives of this study are as follows –

- Comparing above mentioned resampling techniques.
- Assessing the effectiveness of Random Forest on all the resampled datasets.

- Evaluating the model performance using metrics like accuracy, precision, recall, F1-score, AUC-ROC.
- Examining the impact of both the resampling techniques on the model.
- Provide insights into the use of ML techniques for Medicare system.

This paper intends to improve the dependability of resampling in detection of healthcare fraud. Better detection accuracy will help in limiting the financial losses and correct distribution of health-care resources. Moreover, the outcomes could further extend the Machine Learning body of knowledge given that the paper offers information on how to manage imbalanced data, which are also present in numerous industries aside from health care.

Section 2 examines existing methods, identifying weaknesses and the potential of machine learning (ML) to enhance detection. Section 3 will outline the research design, focusing on datasets, resampling techniques, and Machine Learning models used.
Section 4 will be detailing the procedures, outcomes and comparisons between resampling techniques and ensemble models, demonstrating the effectiveness of the proposed approach. Further, in section 5 the results are evaluated based on the metrics and the efficiency of these strategies in addressing healthcare fraud is being analyzed. Finally, section 6 summarizes the key findings, highlight research implications, and offers recommendations for future improvements in fraud detection methodologies, emphasizing the relevance of this research in advancing healthcare fraud prevention.

2 Related Work

Healthcare fraud detection has become a significant research area due to the financial and reputational risks posed to healthcare institutions worldwide. Traditional methods of fraud detection, primarily based on rule-based systems, have been found inadequate to keep up with the increasing complexity of fraud schemes. Previous research works by (Bauder, 2018) (Hancock, 2022) have focused on the application of machine learning techniques for detecting fraudulent transactions in Medicare dataset. In recent years, researchers have also focused on addressing the challenge of imbalanced datasets that commonly arise in healthcare fraud detection.

2.1 Class Imbalance Analysis

Class imbalance in a dataset regards the number of non-fraudulent claims significantly outweighing the number of fraudulent claims, making it a challenge for models to accurately detect the minority class. The issue of class imbalance in Medicare dataset has been a longstanding challenge. In most datasets, fraudulent claims represent a very small proportion of the total claims, making it difficult for machine learning models to learn effective fraud detection patterns. Without addressing this imbalance, models tend to favor the majority class (non-fraudulent claims), leading to suboptimal detection of fraudulent activities. A widely adopted approach to mitigating this problem is the use of resampling techniques (Bounab, 2024).Among the most popular methods are the Synthetic Minority Over-sampling Technique (SMOTE) and its variants.

A recent study by (Bounab, 2024) introduced the SMOTE-ENN hybrid resampling method, which combines Synthetic Minority Over-sampling Technique (SMOTE) with Edited Nearest Neighbors (ENN) to enhance the quality of the minority class data. While SMOTE generates synthetic samples to balance the dataset, ENN eliminates noisy or misclassified instances, thus improving the performance of fraud detection models. The study compared six ML techniques, resulting in Decision Trees (DT) performing the best with the score of 0.99 across all metrics.

A research by (Kennedy, 2024), focused on using unsupervised learning techniques such as autoencoders, to address the challenges pertaining unlabeled variable stating fraud. They have applied the autoencoders on Medicare part D dataset which focuses on identifying anomalies by learning the distribution of majority class and flagging deviations as potential fraud. To analyze the impact of class imbalance on the synthesized class labels, the authors had generated three additionally smaller subsets using Random Under Sampling (RUS) with a class imbalance of 1%, 5% and 20% respectively. These datasets where then trained on by a supervised classifier like DT and compared to the baseline unsupervised Isolation Forest (IF) model, resulting in DT outperforming IF when measured with AUPRC score for all the datasets.

In another recent study, (Leevy, 2023) evaluated the performance of binary classification techniques versus one-class classification (OCC) methods using the Medicare Part D dataset. The study explored algorithms such as CatBoost, XGBoost, and One-Class SVM. The results indicated that binary classification techniques outperformed OCC methods. Among the binary classifiers, decision-tree-based models such as CatBoost delivered the best results scoring 96.93% in AUC and 81.24% in AUPRC.

2.2 Ensemble Methods for Fraud Detection

Ensemble learning methods have gained popularity in recent years for their ability to improve the performance of machine learning models by combining the predictions of multiple models. In the context of healthcare fraud detection, ensemble methods such as Random Forest (RF) and Gradient Boosted Decision Trees (GBDT) have been extensively used by (Hancock, 2022). These methods are particularly effective in handling large and complex datasets, which are characteristic of healthcare fraud detection problems.

Research has shown that addressing class imbalance is critical to improving model performance in fraud detection. In a study by (Cheah, 2023), a hybrid method combining SMOTE and Generative Adversarial Networks (GANs) was proposed with additional integration of FNN and CNN. The use of GANs for resampling allowed the generation of more realistic synthetic fraudulent cases, improving the robustness of the fraud detection models. The authors highlight the potency of GANified-SMOTE when coupled with FNN+CNN classifier provides a better F1-score of 89% for fraudulent data.

A recent study by (Hancock, 2022) evaluated the use of ensemble methods like XGBoost and Random Forest in fraud detection with a focus on maximum tree depth using a highly imbalanced Medicare Part D dataset. The outcomes showed that increasing maximum tree

depth results in increased AUC scores from 0.75 to 0.97 for XGBoost and 0.80 to 0.96 for Random Forest. Another study by Hancock et al. studied the performance of Random Undersampling (RUS) when applied to ensemble classifiers like XGBoost and Extremely Randomized Trees (ET). The authors found that RUS to the 1:9 and 1:27 class ratios, combined with ensemble methods yields the best performance with the AUCROC scores of over 97% (Hancock, 2022).

Moreover, the use of boosting techniques such as AdaBoost and LightGBM has also been explored in healthcare fraud detection. Boosting techniques iteratively improve the model by focusing on errors made by previous models, making them particularly useful in imbalanced datasets. A comparative study by (Cheah, 2023)found that boosting methods like XGBoost, Adaboost and LGBM classifiers showed high accuracies but low F1-scores, Precision, Recall and AUC values suggesting the limited ability of the models to predict minority class. But when combined with SMOTE-ENN showed remarkable results. The study highlighted the importance of hyperparameter tuning in achieving optimal performance with boosting methods, as well as the need for robust cross-validation techniques to avoid overfitting.

2.3 Data Sampling Approaches

Various data sampling approaches like SMOTE, Random Oversampling (ROS), Random Under sampling (RUS), and other hybrid methods, etc have been explored in throughout the years to address the challenge of class imbalance in healthcare fraud detection. One of the most commonly used oversampling methods is SMOTE, which generates synthetic samples for the minority class to balance the dataset. However, SMOTE alone may not always yield the best results, as it can introduce noise into the dataset.

Recent research has focused on hybrid methods that combine oversampling with undersampling to achieve a more balanced and cleaner dataset. In a study by [Bounab et al. (2024)], the authors proposed a hybrid method that combines SMOTE with Edited Nearest Neighbors (ENN). This approach was found to be highly effective in reducing noise in the dataset while simultaneously addressing the class imbalance problem. Similarly, (Hancock, 2022) explored the use of Random Undersampling (RUS) in conjunction with ensemble methods like Random Forest, Decision Tree, XGBoost, etc. The study concluded that RUS, when applied to large datasets, improved the performance of ensemble classifiers by reducing the computational complexity and enhancing model interpretability.

Another promising approach is the use of adaptive synthetic sampling methods such as ADASYN. ADASYN, like SMOTE, generates synthetic samples for the minority class, but it focuses on generating more samples for instances that are harder to classify. This approach has shown to improve model performance in highly imbalanced datasets by creating a more representative and diverse set of synthetic samples. In a study by (Leevy, 2023), ADASYN was compared with SMOTE and other resampling methods, and the results indicated that ADASYN outperformed SMOTE in terms of precision and recall, particularly in fraud detection tasks where the minority class is highly underrepresented.

2.4 Conclusion

In summary, the literature on healthcare fraud detection has evolved significantly in recent years, with a growing emphasis on the use of machine learning techniques to tackle the problem of class imbalance. Resampling methods such as SMOTE, SMOTE-ENN, and SMOTE-GAN have been widely adopted by researchers to address the challenge of class imbalance, while ensemble methods like Random Forest, GBDT, XGBoost, etc have proven effective in handling large and complex datasets. The combination of resampling techniques with ensemble learning offers a powerful solution for improving the accuracy and reliability of fraud detection models in healthcare settings. Future research should continue to explore the potential of hybrid methods and GPU-accelerated models in real-time fraud detection, as well as the application of these techniques to other industries facing similar challenges.

3 Research Methodology

In this section, we discuss the collection and preprocessing of the dataset. We test and evaluate the resampling techniques such as SMOTE-ENN and SMOTE-Tomek that are employed on datasets with varying levels of class imbalance. Further, the model implementation and process on these balanced datasets will be discussed. This section uses the CRISP-DM methodology. Given below in figure 1 is the flow of the research methodology represented by an architectural diagram.

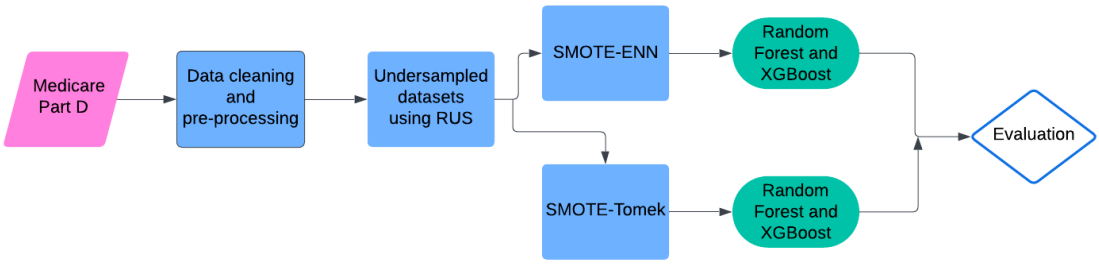


FIGURE 1: ARCHITECTURAL DIAGRAM

3.1 Dataset Preparation and Preprocessing

The dataset utilized in this study consists of Medicare Part D claims data from the latest year 2022 available to download from the CMS (Centers for Medicare & Medicaid Services) website, which includes a variety of healthcare claims, both legitimate and fraudulent (Anon., 2022). The CMS website offers information, tools, and data on U.S. healthcare programs like Medicare and Medicaid for beneficiaries, providers, and researchers. Medicare Part D is a government program in the United States that helps people with Medicare pay for prescription drugs. Medicare Part D also tracks information on prescriptions and spending, which helps to monitor drug use, understand healthcare costs, and identify any fraud.

The Part D data has features which includes details specific to the providers like National Provider Identifier (NPI), medical specialty, gender, geographical information, and claims-level data like the number of beneficiaries per drug, cost, etc. This dataset consists of a total of 22 attributes and approximately 25 million records. To explore the dataset further, the datatypes of each attribute and descriptive statistics have been obtained. Part D data has eleven numerical features and eleven categorical features.

However, it is observed that the Part D data doesn't have a labelled variable stating the fraudulent behaviour. Hence, the data is labelled using LEIE dataset obtained from the United States Office of the Inspector General (OIG) official website (Anon., n.d.). The LEIE dataset is updated monthly by the OIG. It contains information on providers that are excluded from participating in Medicare and Medicaid programs due to fraud convictions. Providers are removed from the LEIE list once the exclusion period is over. As the LEIE data has the NPIs of the providers and exclusion date which has been converted to year format, we join the Part D dataset it by aggregating data over NPIs of the provider. The dataset formed is called as merged df.

With the merged df containing over 25 million records, the research focuses on detecting fraud for a particular state such as Mississippi (MS) to reduce the computational and storage requirements. This dataset is called as MS_df. The state has been chosen randomly for the research. The dataset is searched for any missing or duplicate values. All the seven variables starting with GE65 were dropped as more than 40% of the values were missing, while in the exclusion end year column the missing values are filled by zero value. We filter for those records in the MS df that have exclusion types for providers participating within a year prior to their exclusion end year and label them as fraudulent (Leevy, 2023). The dataset has a binary label where a value of 1 indicates a provider who committed fraud and a value of 0 indicates a provider who has not committed fraud. A correlation matrix is obtained to identify the associations between variables. The variables with high correlation values had been dropped to prevent overfitting, reduce multicollinearity and improve the model performance.

The categorical variables like Prscrbr_City, Prscrbr_Type, Brnd_Name were encoded using the Label Encoder from sklearn library. Next, the Chi-Square test is applied using the 'SelectKBest' method to identify the most influential features for fraud detection. We drop the variable which is least influential. Our processed dataset consists of 270,032 instances and 7 features. The processed data will henceforth be named as 'Data' and the first five columns can be seen in Figure 2. This dataset is highly imbalanced, with only 193 instances (0.0715%) identified as fraud related and 269,839 (99.928%) non-fraud instances. Figure 3 displays a percentage bar chart of fraudulent vs. non-fraudulent providers using a log scale along the y-axis to highlight the very high class imbalance. It gives us a clear view of the highly imbalanced dataset.

	Prscrbr_City	Prscrbr_Type	Tot_Clms	Tot_Day_Suply	Tot_Drug_Cst	Fraudulent
0	166	12	13	28	19.38	0
1	166	12	21	148	113.40	0
2	113	61	20	141	135.65	0
3	113	61	95	708	294.65	0
4	113	61	45	73	64.21	0

FIGURE 2: FIRST FIVE RECORDS OF 'Data'

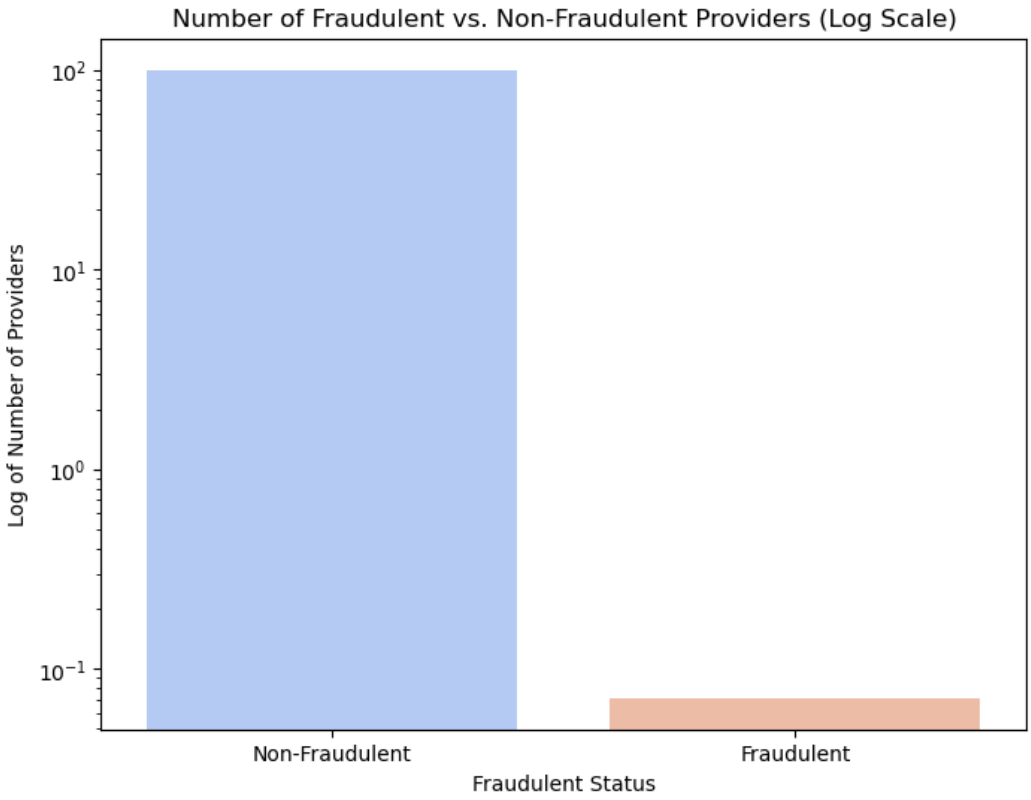


FIGURE 3: PERCENTAGE DISTRIBUTION

Given below in Figure 4 is the descriptive statistics of the Data which provides the overview of the key features like Tot_Clms, Tot_Day_Suply, Tot_Drug_Cst, and the Fraudulent. It includes the total count, mean, standard deviation, minimum and maximum values. The skewness in Tot_Drug_Cst (with a high maximum value) and the class imbalance (fraudulent cases being only 0.07%) are notable. This highlights the dataset's diversity and imbalance challenges, essential for fraud detection modeling.

	Prscrbr_City	Prscrbr_Type	Tot_Clms	Tot_Day_Suply	Tot_Drug_Cst	Fraudulent
count	270032.000000	270032.000000	270032.000000	270032.000000	2.700320e+05	270032.000000
mean	116.823843	33.246737	56.241112	2639.820718	6.698801e+03	0.000715
std	61.664773	14.082725	108.878928	5054.364486	5.051820e+04	0.026725
min	0.000000	0.000000	11.000000	11.000000	0.000000e+00	0.000000
25%	74.000000	18.000000	15.000000	480.000000	2.259375e+02	0.000000
50%	104.000000	31.000000	26.000000	1076.000000	5.770300e+02	0.000000
75%	163.000000	43.000000	55.000000	2519.000000	1.899817e+03	0.000000
max	238.000000	77.000000	17630.000000	146429.000000	1.314515e+07	1.000000

FIGURE 4: DESCRIPTIVE STATISTICS OF PROCESSED DATA

3.2 Handling Class Imbalance

Given the skewed distribution of the classes, with far fewer fraudulent claims (0.0715%) than non-fraudulent ones, balancing the dataset was crucial as the machine learning models can be biased toward predicting the majority class (non-fraudulent claims).
To effectively examine the effects of class imbalance, we generated three additional, smaller sub datasets of the processed data with varying level of class imbalance. To avoid the issue of data leakage, the original dataset is split using the 'traintestsplit()' function resulting in the ratio of 70% training set and 30% testing set. Random Under Sampler from the imblearn library was applied on the train set to randomly under sample from the majority class. RUS is a widely used and effective sampling technique to reduce the class imbalance in a given dataset. We apply RUS to generate three additional datasets with a class imbalance of 1%, 5%, and 20% minority, namely, RUS-1, RUS-5, and RUS-20, respectively (Hancock, 2022). The dataset class characteristics for data as well as the under sampled sub datasets as shown in Table 1.

Dataset	Minority Count	Majority Count	Total Count	Minority Imbalance
Part D				
Full	193	269839	270032	0.0715%
RUS-1	193	19300	19493	0.9901%
RUS-5	193	3860	4053	4.7619%
RUS-20	193	965	1158	16.6667%

TABLE 1: RUS DATASET CLASS CHARACTERISTICS

In addition, we apply the hybrid resampling methods such as SMOTE-Tomek and SMOTE-ENN on the training datasets by defining a function called balance_dataset() to balance these undersampled datasets. The test dataset remains untouched and is used to make predictions. SMOTE technique generates synthetic samples for the minority class while Tomek links removes borderline majority class instances and Edited Nearest Neighbors (ENN) remove noisy or overlapping samples. The parameters for SMOTE-ENN, k_neighbors and n_neighbors have been tuned to obtain better balanced datasets. These hybrid techniques provided cleaner datasets and reduced imbalance, which was expected to enhance model performance. The balanced datasets are saved as a csv file, and the class distribution is displayed in table 3 (ref. section 5).

3.3 Model Development and Training

The Random Forest classifier and XGBoost classifier were selected for this study due to its robustness and ability to handle big data. Random Forest and Extreme Gradient Boost (XGBoost) are both machine learning methods that are based on decision trees to make predictions. Random Forest combines the results of these trees to improve accuracy and avoid overfitting whereas XGBoost builds decision trees one at a time and focuses on fixing errors made by the previous trees.

The models were trained and tested to predict fraudulent activity based on features such as 'Prscrbr_City', 'Prscrbr_Type', 'Exclusion_End_Year', Tot_Drug_Cst and 'Tot_Clms', while 'Fraudulent' as the target variable for all the datasets. Initially, each of the datasets were

split into training and test sets using a train_test_split function in a 70:30 ratio. This split ensured that the model is evaluated on unseen data, providing an unbiased estimate of its performance. Both the models – Random Forest classifier and XGBoost models were trained on the 70% train data for each of the balanced dataset. While training the models, in Random Forest model used balanced class weights and XGBoost Classifier used scale_pos_weight to handle the class imbalance. After training, predictions were made using the 30% test data.

Further the models were evaluated with performance metrics such as accuracy, precision, recall, F1-score, classification matrix and confusion matrix. The confusion matrix for all the datasets were obtained that provided insights into the distribution of true positives, false positives, true negatives, and false negatives, highlighting the model's strengths and weaknesses. The results for each datasets in respect to the models were stored in dictionaries like rf_results and xgb_results. It includes the trained model and evaluation metrics. Finally we print the detailed evaluation results for each dataset, to help analyze the performance across different resampled datasets. AUC-ROC curve and Precision-Recall curve (AUPRC) were also plotted to assess the performance of the model in distinguishing fraudulent from the non-fraudulent cases.

3.4 Model Performance Comparison

The final step involved comparing the performance of the models trained on different resampled datasets. AUC-ROC curves were plotted for each model to visualize their ability to distinguish between the classes. The area under the curve (AUC) provided a single metric summarizing the model's performance across all classification thresholds. Models with higher AUC values were considered better at differentiating between fraudulent and non-fraudulent claims.

In addition to AUC, precision, recall, and F1-scores were compared. Precision measures the proportion of true positive predictions among all positive predictions, while recall measures the proportion of true positives among all actual positives. The F1-score, which is the harmonic mean of precision and recall, provided a balanced measure of the model's performance, especially important in the context of imbalanced datasets.

Overall, the methodology followed a systematic approach to data preprocessing, model training, hyperparameter tuning, and evaluation. This approach ensured that the final model was both effective and generalizable, capable of detecting fraudulent claims in healthcare prescription data.

4 Implementation

Table 2 represents the results of the Chi-Square test, which is a test used in feature selection. It ranks features based on their importance for predicting fraud. It can be observed that Tot_Drug_Cst variable has the most feature significance with a score of 548097. Other notable features that can be seen are Tot_Day_Suply and Prscrbr_City

showing their contribution in predicting fraud. We can also observe that features like Brnd_Name have least relation with the target variable and has been dropped.

	Feature	Chi2 Score
5	Tot_Drug_Cst	548097.91
4	Tot_Day_Suply	85005.084
0	Prscrbr_City	2101.479
3	Tot_Clms	1382.5789
1	Prscrbr_Type	1193.7223
2	Brnd_Name	0.186846

TABLE 2: CHI-SQUARE SCORES

A correlation matrix has been developed using heatmap as shown in figure 5, to find multicollinearity and visualize dependencies between the variables. The scale of the heatmap ranges from -1 (negative correlation) to 1 (positive correlation). The variables like Prscrbr_NPI, Prscrbr_State_Abrvtn, Prscrbr_State_FIPS, Year were dropped due to their high correlation values or they indicated redundancy.
The box plot in figure 6, shows the distribution of total drug cost variable which helps us identify the outliers. This plot helps us detect suspicious costs that have deviated far beyond the expected range, which could indicate fraudulent claims.

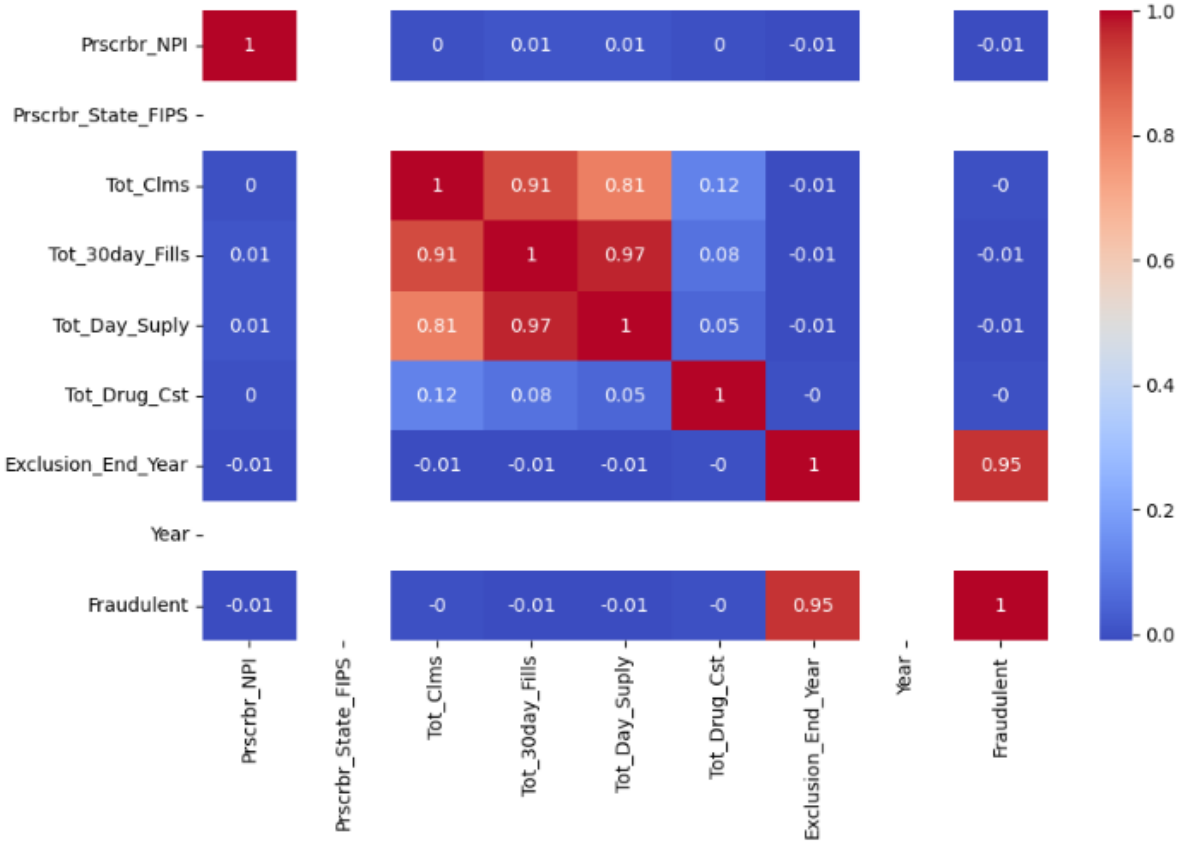


FIGURE 5: CORRELATION MATRIX

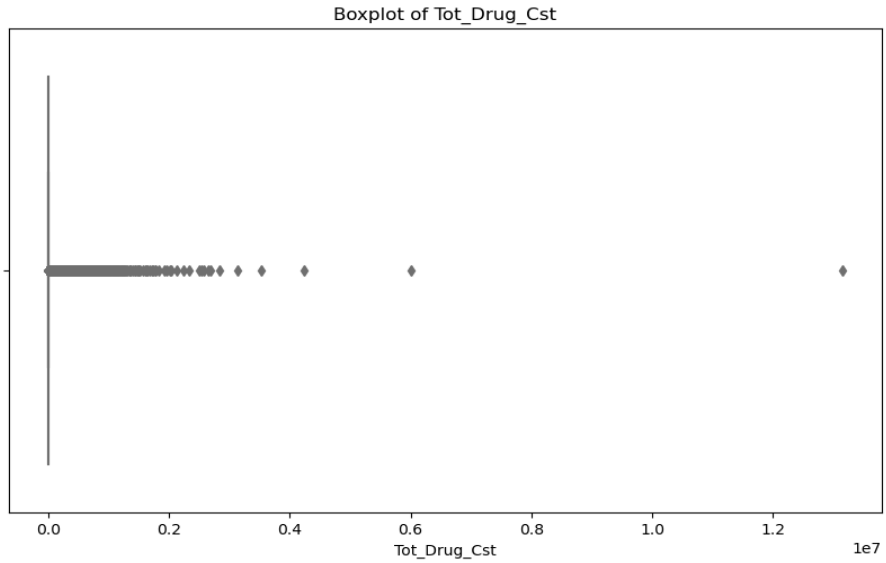


FIGURE 6: BOXPLOT OF TOTAL DRUG COST

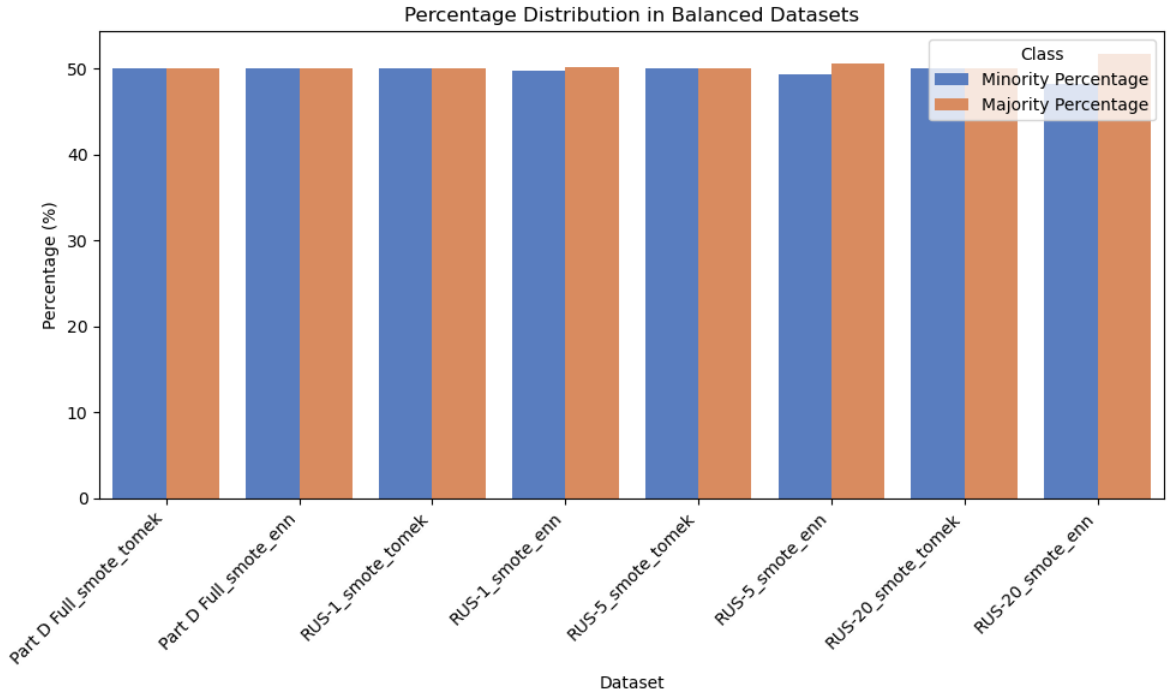


FIGURE 7: DISTRIBUTION IN BALANCED DATASETS

The percentage distribution of the minority and majority classes for all the balanced datasets using SMOTE-Tomek and SMOTE-ENN sampling techniques is displayed in the above figure 7. It can be observed that SMOTE-Tomek achieves perfect balancing whereas SMOTE-ENN shows slight imbalances, particularly for RUS-20_smote_enn dataset. The slight difference in the data balanced using SMOTE-ENN occurs due to its improved elimination of

overlapping and unclear samples. Most of the bar heights shows continuous alignment which indicates the overall effective balance if the classes.

5 Evaluation And Discussion

In this section, we evaluate the performance of all the methods – resampling techniques (SMOTE-Tomek and SMOTE-ENN) and machine learning models (Random Forest and XGBoost), that have been employed in this research for detecting fraudulent claims in Medicare datasets. This analysis was aimed to compare the performance of both the resampling techniques as well as their impact on machine learning models like Random Forest Classifiers and Extreme Gradient Boosting. The evaluation consisted of both, quantitative analysis as well as visualizations to provide an in depth understanding of the model performance.

5.1 Quantitative Results of Resampling Techniques

Dataset	Method	Size	Non-Fraud	Fraud
Part D Full	SMOTE-Tomek	(377774, 9)	188887	188887
Part D Full	SMOTE-ENN	(377774, 9)	188887	188887
RUS-1	SMOTE-Tomek	(27000, 9)	13500	13500
RUS-1	SMOTE-ENN	(27000, 9)	13500	13500
RUS-5	SMOTE-Tomek	(5400, 9)	2700	2700
RUS-5	SMOTE-ENN	(5388, 9)	2700	2688
RUS-20	SMOTE-Tomek	(1350, 9)	675	675
RUS-20	SMOTE-ENN	(1336, 9)	675	661

TABLE 3: BALANCED DATASET CLASS CHARATERISTICS

The class characteristics for balanced datasets represented in table 2, shows the effectiveness of SMOTE-Tomek and SMOTE-ENN in handling class imbalances. For the Part D Full dataset, both the methods achieved near-perfect balance with SMOTE-ENN generating slightly more accurate class distribution by cleaning noisy samples. Whereas in the undersampled datasets like RUS-1 AND RUS-5 and RUS-20, SMOTE-Tomek achieved exact 1:1 class ratio while SMOTE-ENN produced small differences due to aggressive noise reduction. This results the robustness of SMOTE-ENN in handling noise and SMOTE-Tomek's in maintaining precise class imbalance.

5.2 Quantitative Results of Machine Learning Models

Dataset	Accuracy	Precision (0)	Recall (0)	F1-Score (0)	Precision (1)	Recall (1)	F1-Score (1)
Part D Full (SMOTE-Tomek)	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Part D Full (SMOTE-ENN)	1.00	1.00	1.00	1.00	1.00	1.00	1.00
RUS-1 (SMOTE-Tomek)	1.00	1.00	1.00	1.00	0.94	1.00	0.97
RUS-1 (SMOTE-ENN)	1.00	1.00	1.00	1.00	0.95	1.00	0.97
RUS-5 (SMOTE-Tomek)	0.9999	1.00	1.00	1.00	0.92	0.98	0.95
RUS-5 (SMOTE-ENN)	0.9999	1.00	1.00	1.00	0.89	1.00	0.94
RUS-20 (SMOTE-Tomek)	0.9999	1.00	1.00	1.00	0.87	1.00	0.93
RUS-20 (SMOTE-ENN)	0.9999	1.00	1.00	1.00	0.87	1.00	0.93

TABLE 4: RESULTS OF RANDOM FOREST CLASSIFIER

The results for the performance of Random Forest Classifier on the balanced datasets are highlighted in table 4. On the Part D Full dataset, Random Forest Classifier achieved near-perfect accuracy of 99% with both the resampling technique with SMOTE-ENN slightly outperforming SMOTE-Tomek by achieving perfect metrics for all classes. For RUS-1, SMOTE-ENN had higher recall and F1-score for fraudulent claims, whereas SMOTE-Tomek had higher precision showing its strength in reducing false positives. As the size of the dataset decreases, the performance for both the resampling techniques also declines with SMOTE-ENN maintaining better recall.

Dataset	Accuracy	Precision (0)	Recall (0)	F1-Score (0)	Precision (1)	Recall (1)	F1-Score (1)
Part D Full (SMOTE-Tomek)	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Part D Full (SMOTE-ENN)	1.00	1.00	1.00	1.00	1.00	1.00	1.00
RUS-1 (SMOTE-Tomek)	0.9999	1.00	1.00	1.00	0.92	0.98	0.95
RUS-1 (SMOTE-ENN)	0.9999	1.00	1.00	1.00	0.92	0.98	0.95
RUS-5 (SMOTE-Tomek)	0.9999	1.00	1.00	1.00	0.87	1.00	0.93
RUS-5 (SMOTE-ENN)	0.9999	1.00	1.00	1.00	0.87	1.00	0.93
RUS-20 (SMOTE-Tomek)	0.9999	1.00	1.00	1.00	0.87	1.00	0.93
RUS-20 (SMOTE-ENN)	0.9999	1.00	1.00	1.00	0.87	1.00	0.93

TABLE 5: RESULTS OF XGBOOST CLASSIFIER

Table 5 represents the results of XGBoost Classifier and its performance across all the datasets and resampling techniques. It is observed that XGBoost with SMOTE-ENN performed better than XGBoost combined with SMOTE-Tomek for the Part D Full dataset by scoring higher recall and F1- scores for both the classes. On smaller datasets like RUS-1 and RUS-5, SMOTE-ENN retained its advantage in accuracy (98% and 97% respectively), recall and F1-scores but SMOTE-Tomek displayed better precision. However, for RUS-20, the performance declines for both techniques with SMOTE-Tomek performing better compared to SMOTE-ENN by showing 96% accuracy.

5.3 Visual Results of Machine Learning Models

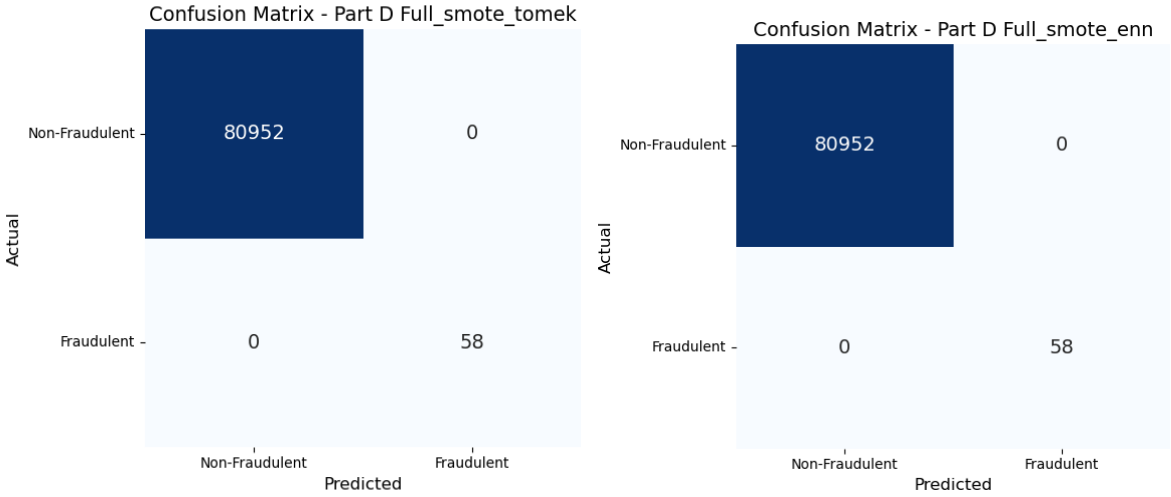
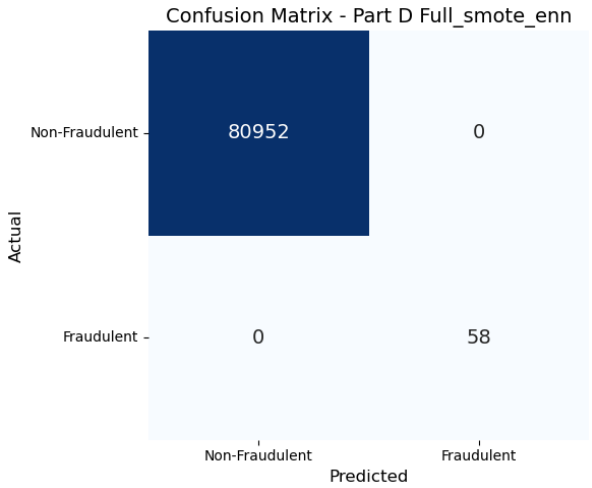


FIGURE 8: CONFUSION MATRIX FOR PART D FULL USING RANDOM FOREST

Figure 8 and 9 represents the confusion matrices for Part D Full dataset using Random Forest Classifier and XGBOOST classifier for both SMOTE-Tomek and SMOTE-ENN



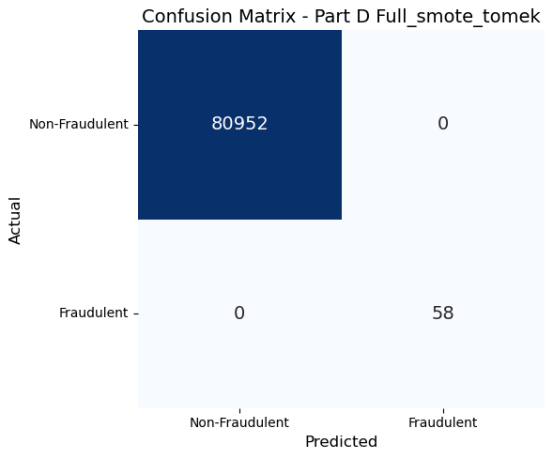


FIGURE 9: CONFUSION MATRIX FOR PART D FULL USING XGBOOST

techniques. On comparing all four confusion matrices, we can say that Part D Full achieves better results for both the models with SMOTE-ENN as both they identified higher number of fraudulent cases correctly and lower number of fraudulent cases were misclassified. Also the Random forest classifier shows slightly better overall performance for both the sampling techniques in compared to XGBoost as fewer non-fraudulent cases are misclassified as fraudulent and vice versa. However, the differences are relatively small and both the models achieve high accuracy in identifying both fraud and non-fraud cases correctly.

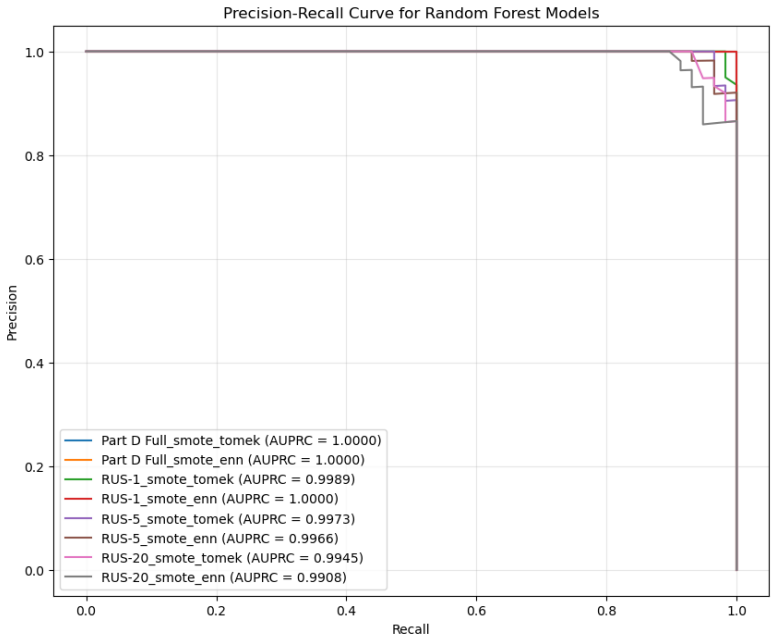


FIGURE 10: AUPRC FOR RANDOM FOREST

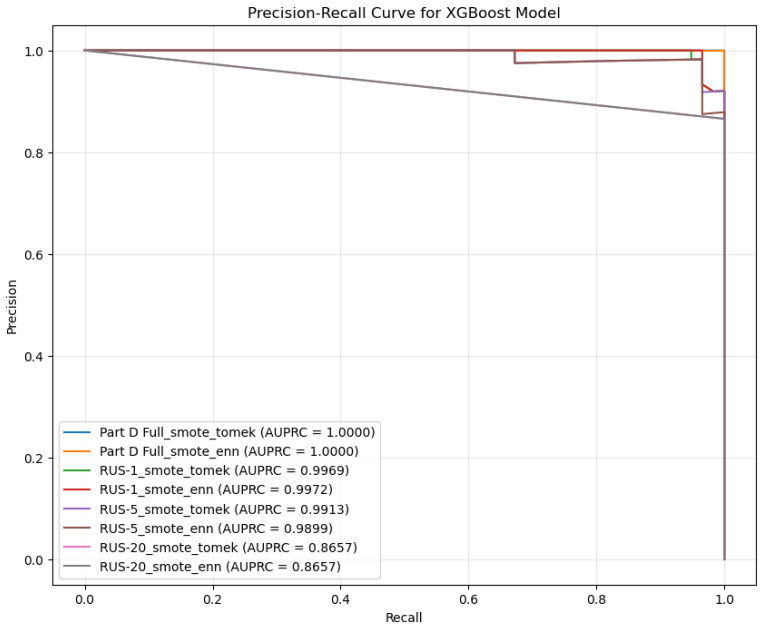


FIGURE 11: AUPRC FOR XGBOOST

Figure 10 displays Precision-Recall Curve (AUPRC) for Random Forest model, whereas figure 11 represents AUPRC Curve for XGBoost models respectively. In both the curves displayed above, the Part D Full_smote_enn achieves the best model performance with the highest Precision-Recall and AUC-ROC values across most of the range. The Part D Full_smote_tomek also has a good performance but not as strongly as the Part D Full_smote_enn for both the Machine Learning models.

6 Conclusion and Future Work

This research investigated the effectiveness of hybrid resampling techniques such as SMOTE-Tomek and SMOTE-ENN to improve the performance of the chosen machine learning model – Random Forest Classifier and XGBoost Classifier, in detecting fraud on the imabalanced Medicare dataset. The class imbalance problem was handled by employing these hybrid resampling techniques on the undersampled datasets and then trained on by the models mentioned above resulting in significant improvements in the model performance.

The results of this study concluded that the hybrid resampling technique- SMOTE-ENN performs better than SMOTE-Tomek with Random Forest achieving better results than Extreme Gradient Boosting model.

Despite achieving superior results, the study is focused on only the Medicare Part D data. Future research could explore these methods on other healthcare datasets or fraud detection domains. Further work could also explore the combination of other hybrid resampling techniques like ADASYN, to handle class imbalance.

The results of this study contribute to the academic literature by demonstrating the real-world advantages of hybrid resampling techniques in fraud detection.

7 References

Anon., 2022. *Center for Medicare and Medicaid Services (CMS).* [Online]
Available at: https://data.cms.gov/provider-summary-by-type-of-service/medicare-part-d-prescribers/medicare-part-d-prescribers-by-provider-and-drug/data
[Accessed 2020].

Anon., n.d. *Office of Inspector General.* [Online]
Available at: https://oig.hhs.gov/exclusions/exclusions_list.asp
[Accessed 2024].

Bauder, R. &. K. T., 2018. *Medicare fraud detection using random forest with class imbalanced big data.* s.l., IEEE International Conference on Information Reuse and Integration (IRI), p. pp. 80–87.

Bauder, R. A. &. K. T. M., 2017. *Medicare fraud detection using machine learning methods.* s.l., 16th IEEE International Conference on Machine Learning and Applications (ICMLA), p. pp. 858–865..

Bauder, R. A. K. T. M. &. H. T., 2018. *Data sampling approaches with severely imbalanced big data for medicare fraud detection.* s.l., IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI), p. pp. 137–142.

Bounab, R. Z. K. G. B. &. K. N., 2024. *Enhancing medicare fraud detection through machine learning: Addressing class imbalance with smote-enn.* s.l., IEEE Access 12, p. 54382–54396.

Cheah, P. Y. Y. &. L. B. ,., 2023. *Enhancing financial fraud detection through addressing class imbalance using hybrid SMOTE-GAN techniques..* s.l., s.n., pp. 11(3), 110..

Hancock, J. &. K. T. M., 2022. *Optimizing ensemble trees for big data healthcare fraud detection.* s.l., IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI), p. (pp. 243–249)..

Hancock, J. e. a., 2023. *Explainable machine learning models for Medicare fraud detection. Journal of Big Data.* s.l., s.n., p. 10(1).

Hancock, J. K. T. M. &. J. J. M., 2022. *The effects of random undersampling for big data Medicare fraud detection.* s.l., IEEE International Conference on Service-Oriented System Engineering (SOSE), p. (pp. 141–146).

J. T. Hancock, R. A. B. H. W. a. T. M. K., 2023. *Explainable machine learning models for medicare fraud detection.* s.l., J. Big Data, p. p. 154.

Kennedy, R. a. K. T., 2024. *Impact of Class Imbalance on Unsupervised Label Generation for Medicare Fraud Detection.* Florida, Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI),, p. pp.53.

Leevy, J. L. H. J. &. K. T. M., 2023. *Assessing one-class and binary classification approaches for identifying Medicare fraud.* s.l., IEEE 24th International Conference on Information Reuse and Integration for Data Science (IRI), p. (pp. 26).

Maas, V. P. a. W., 2013. *Physician medicare fraud: characteristics and consequences.* s.l., International Journal of Pharmaceutical and Healthcare Marketing, pp. vol. 7, no. 1, pp. 8–33,.