IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 March 2024, accepted 2 April 2024, date of publication 8 April 2024, date of current version 22 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3385781

RESEARCH ARTICLE

Enhancing Medicare Fraud Detection Through Machine Learning: Addressing Class Imbalance With SMOTE-ENN

RAYENE BOUNAB[1], KARIM ZAROUR[1], BOUCHRA GUELIB[1], AND NAWRES KHLIFA[2]
[1]LIRE Laboratory, Faculty of New Technologies of Information and Communication, University of Abdelhamid Mehri Constantine 2, Constantine 25016, Algeria
[2]Research Laboratory of Biophysics and Medical Technologies, Higher Institute of Medical Technologies of Tunis, University of Tunis El Manar, Tunis 1006, Tunisia

Corresponding author: Rayene Bounab (rayene.bounab@univ-constantine2.dz)

ABSTRACT The healthcare fraud detection field is constantly evolving and faces significant challenges, particularly when addressing imbalanced data issues. Previous studies mainly focused on traditional machine learning (ML) techniques, often struggling with imbalanced data. This problem arises in various aspects. It includes the risk of overfitting with Random Oversampling (ROS), noise introduction by the Synthetic Minority Oversampling Technique (SMOTE), and potential crucial information loss with Random Undersampling (RUS). Moreover, improving model performance, exploring hybrid resampling techniques, and enhancing evaluation metrics are crucial for achieving higher accuracy with imbalanced datasets. In this paper, we present a novel approach to tackle the issue of imbalanced datasets in healthcare fraud detection, with a specific focus on the Medicare Part B dataset. First, we carefully extract the categorical feature ''Provider Type'' from the dataset. This allows us to generate new, synthetic instances by randomly replicating existing types, thereby increasing the diversity within the minority class. Then, we apply a hybrid resampling method named SMOTE-ENN, which combines the Synthetic Minority Over-sampling Technique (SMOTE) with Edited Nearest Neighbors (ENN). This method aims to balance the dataset by generating synthetic samples and removing noisy data to improve the accuracy of the models. We use six machine learning (ML) models to categorize the instances. When evaluating performance, we rely on common metrics like accuracy, F1 score, recall, precision, and the AUC-ROC curve. We highlight the significance of the Area Under the Precision-Recall Curve (AUPRC) for assessing performance in imbalanced dataset scenarios. The experiments show that Decision Trees (DT) outperformed all the classifiers, achieving a score of 0.99 across all metrics.

INDEX TERMS Healthcare fraud, imbalanced data, machine learning (ML), noisy data.

I. INTRODUCTION

Healthcare systems globally face a significant challenge due to fraud, which impacts both their financial stability and moral principles. In particular, the U.S. Medicare program, a key element of the healthcare sector, experiences substantial financial loss from such fraudulent practices. According to the Federal Bureau of Investigation, healthcare fraud represents 3–10% of the total healthcare costs, leading to yearly losses between $19 billion and $65 billion [1]. These illegal activities not only deplete financial resources but also affect the operational efficiency and trustworthiness of healthcare systems. Therefore, it is imperative to implement effective and strong fraud detection strategies, especially in Medicare, which serves a broad and diverse population. Ensuring efficient fraud detection is vital for the protection of public funds and guaranteeing that resources are distributed fairly for necessary healthcare services and patient care. The challenge in healthcare fraud detection lies in the evolving nature of fraud schemes, which are complex and diverse. Traditional, rule-based detection methods fall short in this dynamic environment, lacking the necessary adaptability and

The associate editor coordinating the review of this manuscript and approving it for publication was Muammar Muhammad Kabir.

scalability to address the sophisticated nature of modern healthcare fraud. Machine learning (ML), a subfield of Artificial intelligence (AI) has demonstrated exceptional proficiency in healthcare fraud detection, particularly in processing the Medicare dataset released annually by the U.S. government [2]. This dataset is a crucial resource for researchers focusing on healthcare fraud detection. This reflects the government's commitment to combating fraud by equipping specialists with vital data, thereby facilitating the development of more sophisticated fraud detection strategies based on ML. Its strength lies in its ability to learn from historical data and adapt to emerging fraudulent patterns, making it effective in analyzing large datasets to identify anomalies and fraud indicators. This adaptability renders ML indispensable in creating efficient, responsive systems for large-scale operations like Medicare, positioning it is an indispensable asset in combating healthcare fraud [3]. Machine Learning (ML), however, excels in this aspect. Its ability to learn from historical data and adjust to new fraudulent patterns allows it to process and analyze vast datasets, detecting anomalies and patterns indicative of fraud. This capability positions ML as a crucial tool in creating more effective and responsive fraud detection systems, especially for large-scale operations like Medicare. Its dynamic approach makes it an indispensable asset in the ongoing fight against healthcare fraud [3].

Recent studies, such as those by [4], [5], [6], [7], and [8] demonstrate the successful application of ML techniques using the Medicare dataset to uncover fraudulent activities. The Medicare datasets [9], disseminated by the Centers for Medicare and Medicaid Services, exhibit a pronounced class imbalance characterized by a disproportionate representation of non-fraudulent cases relative to fraudulent instances. This class imbalance presents a formidable impediment to the efficacy of ML algorithms deployed for fraud detection. Predominantly, ML models are predisposed to a bias towards the majority class, in this case, non-fraudulent transactions, leading to a heightened incidence of false negatives. This phenomenon occurs when the algorithm erroneously categorizes fraudulent activities as legitimate, a direct consequence of the skewed training data [2], [10]. Such imbalance in the dataset precipitates the development of ML models that demonstrate suboptimal performance in the accurate detection of fraudulent activities. This deficiency critically undermines the overarching effectiveness and reliability of the fraud detection mechanism within the healthcare domain. To ameliorate this situation, it is imperative to establish datasets that are balanced, thereby ensuring that ML algorithms are more adept at discerning the minority class, which in this context refers to fraudulent transactions. A balanced dataset is instrumental in enabling the algorithm to detect nuanced patterns and anomalies that are indicative of fraudulent activities [5].

A notable gap in current research endeavors within healthcare fraud detection is the inadequate focus on addressing the challenges posed by imbalanced data. The preponderance of research has been directed towards classification tasks, with insufficient attention to the intricate issue of data imbalance. Although there has been a notable deficiency in addressing data imbalances within healthcare fraud detection, some researchers have begun to address this gap using resampling techniques. These methodologies, which include Random Oversampling (ROS) [5], Adaptive synthetic sampling approach (ADASYN) [11], and Synthetic Minority Over-sampling Technique (SMOTE) [12]. Concurrently, undersampling of the majority class is executed using Random Undersampling (RUS) [13] to achieve a balanced dataset. Despite the efficacy of these techniques, challenges persist. ROS methods, for instance, may be susceptible to overfitting, potentially compromising the generalizability of the model. Meanwhile, the application of SMOTE carries the risk of introducing noise to the dataset. Moreover, the implementation of RUS comes with its own set of concerns, notably the risk of discarding crucial data, potentially leading to a loss of important information. The intricate trade-offs and considerations associated with each resampling technique underscore the complexity of addressing the class imbalance in healthcare fraud detection datasets. To address the limitations identified in prior research, we focus on three main areas:

- Advancing research into techniques for managing imbalanced datasets
- Evaluating resampling approaches, with an emphasis on the drawbacks of ROS, which can cause overfitting, and SMOTE, which may add noise to the dataset.
- Examining the impact of RUS on the potential loss of essential data, which could lead to overlooking critical indicators of fraud.

This paper introduces a novel approach to address imbalanced datasets in healthcare fraud detection, particularly focusing on the Medicare Part B dataset. A key innovation lies in the meticulous separation of the categorical features from the numerical features, enabling the random generation of synthetic instances to enrich minority class diversity. Our proposed Synthetic Minority Over-sampling technique with Edited Nearest Neighbors (SMOTE-ENN) hybrid resampling method contributes significantly by simultaneously rebalancing the dataset and eliminating noisy data, which is then evaluated using various ensemble classifiers. To the best of our knowledge, this paper proposes an approach that combines the separate generation of categorical features, with the SMOTE-ENN technique and a variety of ensemble learning classifiers. Additionally, we incorporate the use of the Area Under the Precision-Recall Curve (AUPRC) metric for evaluation, enhancing the robustness and comprehensiveness of our analysis.

The main contributions of this paper can be summarized as follows:

- Randomly generate the categorical feature "Provider. Type" based on existing categories in the dataset
- Application of the SMOTE-ENN hybrid resampling method to balance the dataset and remove noisy data.

- Evaluation of the effectiveness of the proposed approach using ensemble learning classifiers.
- Employing the Area Under the Precision-Recall Curve (AUPRC) metric for a more effective evaluation of model performance in the context of an imbalanced dataset

The structure of this paper is organized as follows: Section II provides an overview of the related work, with an emphasis on studies that utilized ML and data balancing techniques. The proposed system is detailed in Section IV. The experimental results and a discussion are presented in Section V. Finally, the paper concludes with Section VI, summarizing the main outcomes.

II. RELATED WORK

Detecting fraud in healthcare has been the subject of extensive exploration in the literature. This section presents and evaluates different papers in the field of healthcare fraud detection based on two principal aspects that align with the objectives of our study. Firstly, there is a significant amount of literature that focuses on the utilization of AI methodologies to detect healthcare fraud. Many studies highlight the effectiveness of ML techniques in effectively identifying fraudulent behavior within healthcare systems [14]. Another area of research examines the challenge of imbalanced data in healthcare fraud detection. Researchers have explored various strategies to handle this problem, aiming to improve the effectiveness of ML models in accurately detecting healthcare fraud [15].

A. WORKS ADDRESSING THE USE OF AI IN HEALTHCARE FRAUD DETECTION

Recent advancements in AI, especially ML, have led to diverse and innovative approaches to detecting healthcare fraud. The authors in [16] aimed to improve decision-tree-based ensemble techniques for healthcare fraud detection, utilizing the large Part D Medicare dataset with around 175 million records. The authors in [17] introduced a ML framework that transforms prescription claims into statistical modeling features, focusing on business heuristics, provider-prescriber relationships, and client demographics. The study by [2] employed an ensemble feature selection technique in ML models for Medicare fraud detection. This approach improved explainability and reduced data complexity. The work proposed by [18], introduced a Bayesian Belief Network (BBN) model for healthcare fraud detection, involving preprocessing and feature engineering of Texas Medicaid prescription claims. This approach outperformed baseline models in scalability and interpretability.

In [19], the authors concentrated on applying a data-centric AI approach to detect U.S. Medicare fraud. This significantly enhanced ML models' performance through careful data preparation and feature engineering. Their approach showed superior results compared to traditional datasets in Medicare fraud classification tasks. Reference [6] proposed a study to detect healthcare fraud instances by applying four ML algorithms. In their research, they identified 19 essential features, which they organized into four primary categories.

Upon examining the studies, we can observe the use of diverse methods in detecting fraud, such as ensemble methods, decision-tree-based techniques, and BBN. Moreover, several works emphasize the important role of data preparation, feature engineering, and feature selection in enhancing the model's performance. However, a common limitation observed is the reliance on the significantly imbalanced Medicare dataset for experimentation, an issue that remains largely unaddressed and could potentially result in misclassification outcomes.

B. WORKS ADDRESSING THE PROBLEM OF IMBALANCED DATA IN HEALTHCARE FRAUD DETECTION

The following studies present some of the common methods applied in the field of healthcare fraud detection to handle imbalanced data.

The paper [20] tackled the problem of imbalanced data by experimenting with different class distributions in their ML models. Using the Medicare Part B dataset, the authors applied six ML models across seven class distributions to address the data imbalance. The results indicate that employing a 90:10 ratio of non-fraud to fraud cases outperformed other models. In their study, [21], the authors addressed the challenge of the imbalanced data in the Medicare dataset by employing ML models for classification and six sampling techniques to balance the dataset. The study's findings demonstrated that RUS consistently gave strong results across all ML models. A semantic embedding approach was proposed in [22]. The author proposed a semantic embedding approach to convert healthcare procedure codes (HCPCS) from the Medicare fraud dataset into semantic embeddings. To address the imbalanced data issue, the work employed a simple undersampling method. Another semantic embedding approach was proposed in [23]. The authors developed semantic embeddings for medical provider types using both pre-trained (Global Vectors for Word Representation (GloVe), Medical Word2Vec (Med-W2V)) and custom (HcpcsVec, RxVec) embeddings from Medicare claims data. This method improved the representation of provider specialties and was validated using various ML algorithms. Additionally, the study tackled the issue of imbalanced data by employing random over-sampling (ROS) and under-sampling techniques. The authors in [14] applied ML and DL techniques to identify financial fraud in healthcare credit card transactions. Additionally, they tackled the challenge of imbalanced data by recommending a hybrid resampling approach, although the study did not specify the particular methods used for this resampling.

In their study, [24], the authors proposed unsupervised DL techniques to detect procedure code overutilization in medical claims. To tackle the imbalanced data, the test set was composed of outliers representing potential fraudulent cases. The paper, [25], focused on assessing the performance of ML

classifiers in the Medicare imbalanced dataset. The authors applied the RUS method with various ensemble learning techniques to address class imbalances. Another paper, [26], explored the classification of healthcare fraud using the highly imbalanced Medicare dataset by employing the RUS method to address the imbalance issue. The results show RUS enhanced the AUC scores while reducing the training data size. In the paper [11], the authors proposed the use of two data balancing techniques, namely: Class Weighing Scheme (CWS) and ADASYN. Moreover, to classify instances, the authors applied a range of ML algorithms.

C. DISCUSSION OF RELATED WORK

The reviewed papers demonstrate a focus on employing ML techniques for detecting various forms of healthcare fraud. A significant challenge across these studies is the management of imbalanced datasets, a prevalent issue in fraud detection. References [21], [22], [25], and [26] applied the RUS method, which randomly removed samples from the majority class to match the number of samples in the minority class. While it reduces time complexity and computational load, RUS significantly limits healthcare fraud detection. Its main drawback is the potential loss of critical information, as it randomly removes majority-class instances. On the other hand, ROS, as applied by [23], can be effective in providing a balanced dataset without losing information. However, it can lead to overfitting. By duplicating minority class samples, ROS can make the model specific to the existing fraud instances, reducing its generalizability to new or slightly different types of fraud. Thus, it is important to apply new methods that generate new instances, such as the SMOTE method. One paper in the related works adopted a hybrid resampling approach [14], which combines under-sampling and over-sampling methods to mitigate their drawbacks. However, there is a lack of information regarding the methods used. Their hybrid approach leaves a gap in understanding its efficacy and applicability in diverse healthcare fraud scenarios. Moreover, in [11], two methods were applied, namely: ADASYN and CWS. ADASYN generates synthetic samples for the minority class based on a density distribution, which helps in creating more diverse and representative samples. However, in complex Medicare fraud datasets, this method can introduce noise. Overall, the major gaps in the studies on Medicare fraud detection using ML largely stem from an inadequate exploration of more sophisticated techniques to handle imbalanced datasets. There is a need for methods that can manage the complex, high-dimensional nature of Medicare data. The SMOTE method, known for its effectiveness in generating representative minority class samples, is notably not well explored in this field. Additionally, the problem of noisy data when generating new instances is not discussed; thus, taking this challenge into account is important when dealing with imbalanced datasets. Leveraging the power of hybrid methods should also be taken into account. Furthermore, these resampling methods could be significantly enhanced when combined with ensemble learning classifiers, known for their robustness and generalizability. Addressing these gaps with such advanced methodologies could significantly improve the accuracy and efficiency of healthcare fraud detection in Medicare systems.

Table 1 provides a comprehensive comparative overview of the related works in the field of healthcare fraud detection. It details the datasets used, ML methods applied, data balancing techniques employed, and the evaluation metrics achieved in each study.

III. PROBLEM FORMULATION

Given the significant class imbalance in the Medicare Part B dataset, as illustrated in Figure 1, with a ratio of 1:11,312 between fraudulent (minority class) and non-fraudulent claims (majority class), traditional ML models face substantial challenges in accurately detecting instances of fraud. This imbalance biases models towards the majority class, severely undermining their capability to generalize and identify fraudulent activities effectively.

To address this imbalance, we propose the use of the SMOTE-ENN algorithm. We denote the set of fraud detection models as $\{f_m\}_{m=1}^{M}$, each trained on its respective subset of data $\mathcal{D}_m$, where $\mathcal{D}_m = \{(\mathbf{x}_i^m, y_i^m)\}_{i=1}^{N_m}$. Here, $\mathbf{x}_i^m$ represents the feature vector for the i-th claim, and y_i^m indicates its corresponding class label.

The SMOTE-ENN algorithm (see Algorithm 1) is applied to each subset $\mathcal{D}_m$, to generate a balanced dataset $\mathcal{D}'_m$ through synthetic sample generation and noise reduction. This process is formulated as:

$$\mathcal{D}'_m = \text{SMOTE-ENN}(\mathcal{D}_m). \tag{1}$$

The primary challenge is to validate the effectiveness of the SMOTE-ENN approach to balancing the dataset and improving the detection accuracy of the models f_m. The performance of the models trained on the balanced dataset $\mathcal{D}'_m$ will be assessed and compared to their performance on the original imbalanced dataset $\mathcal{D}_m$, with a focus on their accuracy and generalization in detecting fraudulent activities.

IV. PROPOSED SOLUTION

This section outlines the proposed solution, focusing on achieving dataset balance. It begins by generating categorical data and applying the SMOTE-ENN technique to numerical data, specifically for the classification task. The approach will be detailed in two phases: first, providing a comprehensive overview of the entire architecture, followed by a thorough explanation of each component.

A. OVERALL OVERVIEW

The significant disparity in class distribution shown in the Medicare dataset poses a difficult obstacle to effectively identifying fraudulent claims. This challenge highlights the need for a strong technique that can fix imbalances to enhance model performance. Equation 1 presents the SMOTE-ENN algorithm as our recommended solution, which tackles the

TABLE 1. Comparative table of related works.

Ref	Dataset	ML Methods	Data Balancing Method	Evaluation
[21]	Medicare	Logistic Regression(LR), Random Forest (RF), Gradient Boosting Trees (GBT)	ROS, RUS, SMOTE, SMOTE variants, ADASYN	Area Under the Curve (AUC)= 0.82
[20]	Medicare Part B	Naive Bayes(NB), LR, Decision Trees (DT), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), RF	RUS	AUC
[22]	Medicare	Word2Vec (Skip-gram, Continuous Bag Of Words (CBOW))	Undersampling	AUC=0.870, Geometric Mean (G-mean)=0.783
[23]	Medicare	Logistic Regression (LR), RF, GBT, Multi-Layer Perceptron (MLP)	ROS, RUS	AUC=0.830
[14]	Healthcare Transactions	NB, LR, KNN, RF, Convolutional Neural Network (CNN)	Hybrid Resampling	Accuracy=97.58
[16]	Part D Medicare	eXtreme Gradient Boosting (XGBoost), RF	-	AUC= 0.97
[17]	Prescription Claims	LR, RF, Principal Component Analysis(PCA)	-	Receiver Operating Characteristic (ROC)= 0.76, F1-score= 0.88
[11]	Healthcare insurance	LR, DT, RF, XGBoost	CWS, ADASYN	AUC=0.95
[2]	Medicare	Category Boosting (CatBoost), XGBoost, RF, Extremely Randomized Trees(ET), Light Gradient Boosting Machine (LightGBM), DT, LR, Ensemble Feature Selection	-	AUC= 0.95, Area Under the Precision-Recall Curve (AUPRC)=0.78
[25]	Medicare	CatBoost, XGBoost, LightGBM, RF, ET	RUS	AUC=0.97, AUPRC=0.92
[26]	Medicare	CatBoost, XGBoost, RF, ET	RUS	AUC=0.99
[19]	U.S. Medicare	XGBoost, RF	-	G- mean = 0.90, AUC = 0.962
[18]	Texas Medicaid	Bayesian Belief Network(BNN)	-	F-score=0.94
[6]	Healthcare Insurance	SVM, DT, RF, MLP	-	F-score=0.95
[24]	Healthcare Claims	Deep Autoencoders	-	precision=0.87, recall=1.00, F-score=0.93

imbalance by increasing the size of the minority class and refining the dataset.

Figure 1 presents the proposed architecture for healthcare fraud detection based on generating categorical data and the SMOTE-ENN method formulated in 1. In our architecture for classifying healthcare fraud claims, we first partition the dataset based on data type and then apply a series of preprocessing steps to enhance data quality. To tackle data imbalance, we utilize SMOTE-ENN, a hybrid resampling method. Along with this, we augment categorical data using the Random Sampling without Replacement method. Finally, we employ various ensemble learning classifiers for classification.

B. DATA COLLECTION

The datasets used for the study include publicly accessible Medicare data Physician and Other Practitioners (PartB) of the year 2020, provided by the Centers for Medicare & Medicaid Services (CMS) [27], and the List of Excluded Individuals and Entities (LEIE).

The Medicare dataset was obtained in a comma-delimited format (CSV), making it suitable for additional data processing procedures. To facilitate an in-depth comprehension of the data, the CMS provides methodological documentation that clarifies its techniques for collecting and processing data. This is further supported by data dictionaries that outline the definitions of all attributes present in the datasets. The proposed study is specifically centered on the dataset known as ''Medicare Part B Summary by Provider and Service 2020.'' This dataset contains approximately 9,449,361 records and a range of 29 distinct features. Many of the attributes are provider demographic data, which we do not use for modeling purposes. As a result, it serves as a valuable resource for our analytical investigations. The List of Excluded Individuals and Entities (LEIE) is managed by the Office of Inspector General (OIG) in compliance with Sections 1128 and 1156 of the Social Security Act [28]. The Office of Inspector General (OIG) maintains the authority to exclude healthcare providers from engaging in federally financed healthcare programs due to a range of legitimate reasons. It is worth mentioning that individuals who are placed on the exclusion list are considered unable to receive payments from Federal healthcare programs for any services that they provide. To pursue reinstatement, those who have

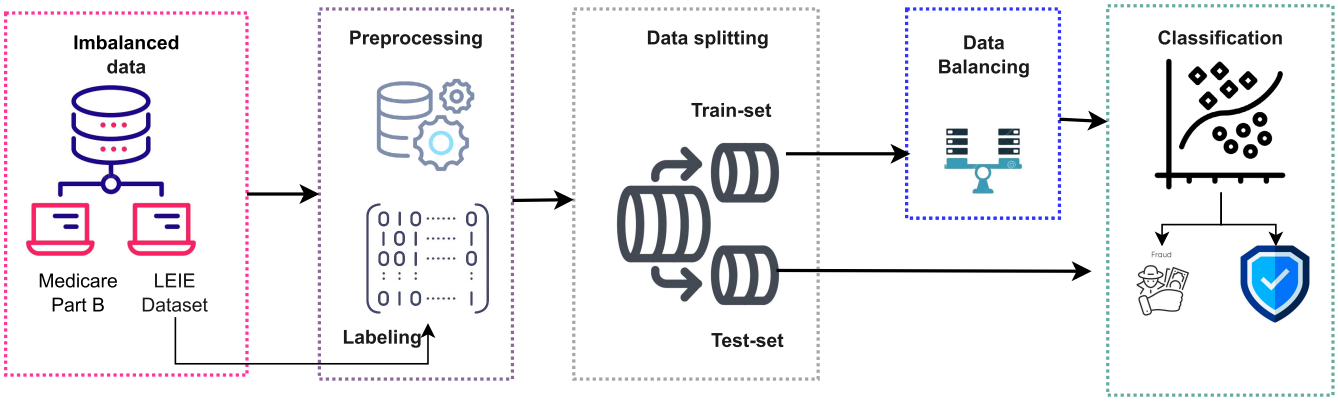

FIGURE 1. Proposed architecture for healthcare fraud detection based on SMOTE-ENN.

been excluded must adhere to a prescribed procedure after successfully fulfilling the duration of their exclusion. The current structure of the LEIE data consists of 18 attributes that give relevant information regarding the provider under investigation and outline the precise reasons for their exclusion.

C. PREPROCESSING

Data were carefully preprocessed under the Centers for Medicare & Medicaid Services (CMS) methodological documentation, which provides valuable insights into their data processing procedures and comprehensive data dictionaries that meticulously define the dataset's attributes. We carefully follow the data preparation method proposed in [19]. We start by adding a new column that serves later for labeling, denoted as "Year." The value "2020" is assigned, representing the year of the dataset. Following this, we move on to identifying and rectifying any instances of missing values. To facilitate this procedure, we utilized the methodology described in the data dictionary files that were supplied by the Centers for Medicare & Medicaid Services (CMS) [29]. The process of imputing missing data was conducted systematically. Specifically, when faced with missing information regarding the gender of providers, we introduced a third category denoted as "U" to represent unknown values. Next, the gender value was encoded numerically, with the assignment of M=1, F=0, and U=2. A comprehensive assessment was conducted to analyze the characteristics of the missing values in the remaining columns, particularly concerning the absence of provider names and geographic details. Due to their low relevance to our study, we have chosen to eliminate these columns from further consideration. Our subsequent step involved the selection of specific rows that met the condition of having the value 'N' in the 'HCPCS_Drug_Ind' column as recommended in the CMS documentation. The second dataset used in this study, which plays an essential role in the labeling procedure of the Part B dataset, is the LEIE dataset. It is formatted as a character-separated value file (CSV). The relevant features of this dataset are NPI, the exclusion type, the exclusion date, waiver data, and the reinstatement date. In the LEIE CSV file, these elements are named: NPI, EXCLTYPE, EXCLDATE, REINDATE, WAIVERDATE, and WVRSTATE, respectively. We followed the same methodology presented in [20], to prepare the LEIE data. After preparing the two datasets, we proceed to the labeling step using the LEIE dataset. During the process of labeling, two crucial criteria are created for the detection of fraudulent activities: First, the National Provider Identifier (NPI) from Part B claims should be present in the LEIE dataset. Secondly, the year of Part B precedes the year in which the exclusion period concludes. When these conditions are met, the record is labeled as fraud; otherwise, it is labeled non-fraud. The labeling technique we describe here is the same technique outlined in [8], [30], [31], and [32]. After labeling the dataset, we remove the columns with low pertinence in the dataset, namely: NPI, YEAR, HCPCS_Drug_Ind, and keep only 9 features. Table 2 presents the features used in experiments based on work [31].

Finally, we normalize the dataset to ensure that each feature contributes equally during analysis or modeling. This phase protects statistical learning methods by preventing larger numeric values from overwhelming smaller ones [33].

D. SPLITTING DATASET BASED ON DATA TYPE

In this step, the dataset is divided according to the data type, whether numerical or categorical, to facilitate separate treatment and preservation of the local structure of the information. The dataset comprises eight numerical features and one categorical feature, denoted as "Rndrng_Prvdr_Type". Figure 2 represents the splitting into numerical and categorical data.

The provider type attribute is a categorical variable that describes the provider or supplier's medical speciality, which encompasses 102 distinct types (e.g., Internal Medicine, Family Medicine, Cardiology, etc.). The objective is to generate instances based on the existing "Provider Types". To accomplish this, the Random Sampling without Replacement method is employed. Initially, the 102 provider types are shuffled to ensure a random order. Subsequently, each Provider type is selected sequentially from this shuffled

TABLE 2. Description of medicare data features.

Feature	Description	Type
Rndrng_Prvdr_Gndr	Provider Gender	Categorical
Rndrng_Prvdr_Type	Type of Provider	Categorical
Tot_Benes	Number of Medicare Beneficiaries	Numerical
Tot_Srvcs	Number of Services	Numerical
Tot_Bene_Day_Srvcs	Number of Distinct Medicare Beneficiary/Per Day Services	Numerical
Avg_Sbmtd_Chrg	Average Submitted Charge Amount	Numerical
Avg_Mdcr_Alowd_Amt	Average Medicare Allowed Amount	Numerical
Avg_Mdcr_Pymt_Amt	Average Medicare Payment Amount	Numerical
Avg_Mdcr_Stdzd_Amt	Average Medicare Standardized Amount	Numerical

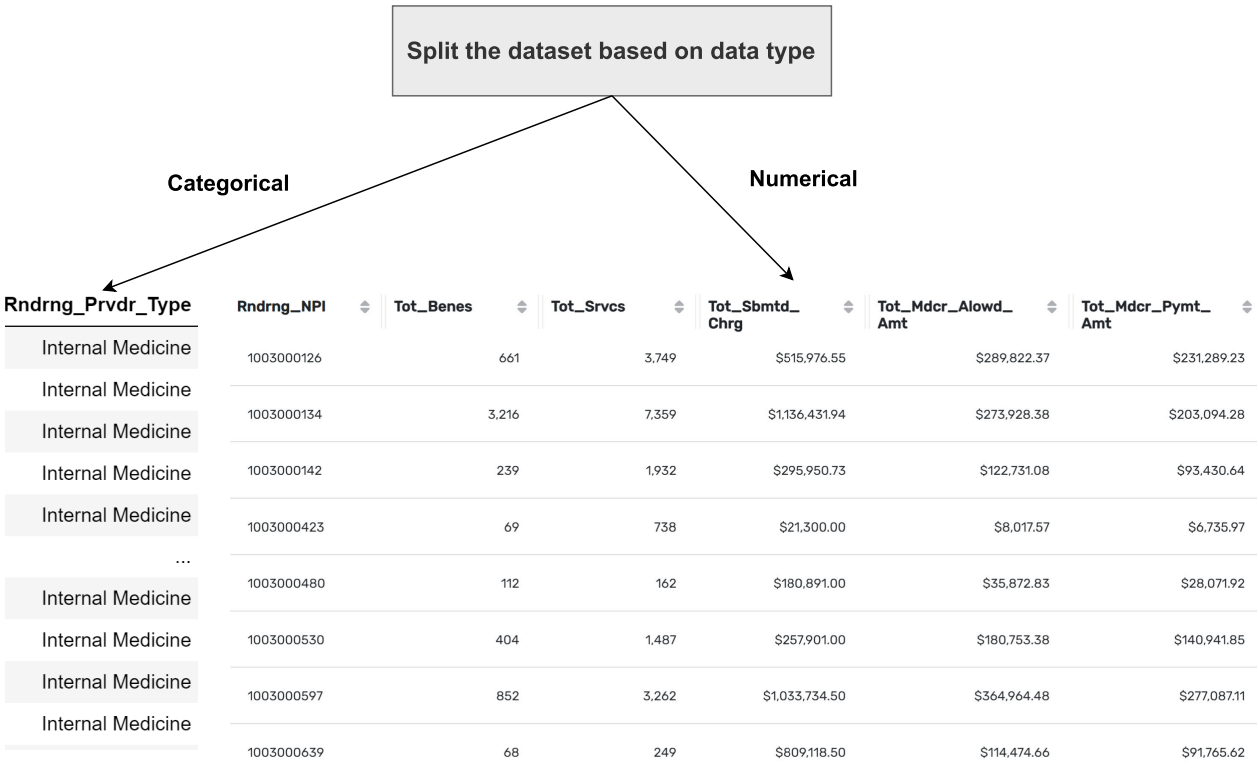

FIGURE 2. Splitting the dataset based on data type.

list, ensuring every type is selected once before any type is selected again. After exhausting all 102 types, the list is reshuffled, and the selection process is repeated. This process continues until the desired number of k instances is reached. This method guarantees that each provider type is represented fairly and equally across the total instances, preventing any bias towards certain types. Figure 3 explains the generation of ''ProviderType'' based on 102 existing types.

E. TRAIN-TEST-SPLIT

To accurately evaluate our models' performance, we divide our dataset into training and test sets, using the ''Train_test_split'' method. This approach enables assessing the models' ability to perform effectively on new, unseen data and determining their overall efficacy. We split the dataset into Train_Test_Sets based on the ratio 80:20, where 80% of the dataset was assigned to the training set, while the remaining 20% constitute the test set.

F. SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE WITH EDITED NEAREST NEIGHBORS (SMOTE-ENN)

SMOTE-ENN is a composite resampling approach, amalgamating the principles of both oversampling and undersampling to tackle the challenge of imbalanced datasets, as proposed by [34]. The initial phase involves augmenting the minority class representation through the *SMOTE* algorithm, which synthesizes synthetic instances by linear interpolation between existing minority class samples and their nearest neighbors [35]. Nevertheless, the randomness in selection intrinsic to *SMOTE* can introduce noise, potentially impeding the model's ability to generalize [36]. To mitigate such effects, the Edited Nearest Neighbor (*ENN*) method is employed after *SMOTE*. This subsequent step aims to purify the dataset by discarding instances that introduce noise or redundancy. The procedure involves examining each instance to ensure the consistency of its class label with those of its nearest neighbors, thus enhancing the dataset's

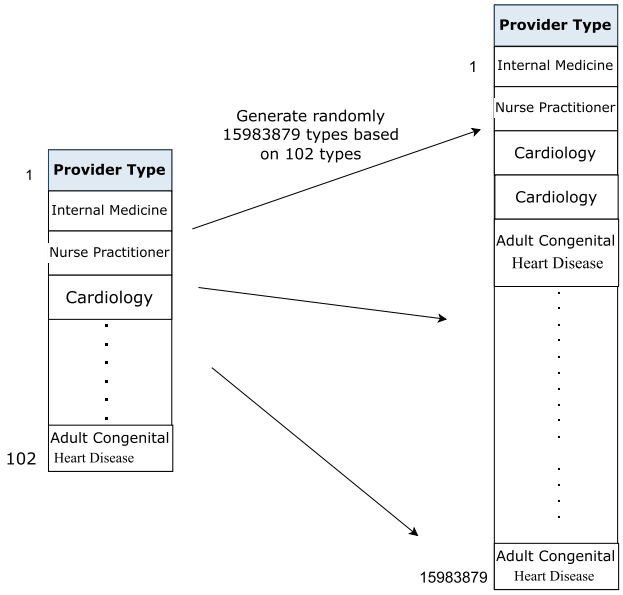

FIGURE 3. Generation of provider type.

overall quality for subsequent modeling. Using the SMOTE-ENN approach, we generated 7,119,172 synthetic instances, thereby balancing the dataset and achieving a better class distribution.

Figure 4 demonstrates the detailed steps of this procedure, showing how SMOTE generates synthetic instances and ENN eliminates noise examples. Having a strong methodological foundation is essential for creating reliable fraud detection algorithms that can generalize effectively across various types of claims. The relationship between the problem statement presented in Equation 1 and the methodological framework shown in Figures 1 and 4 demonstrates our thorough strategy for addressing the class imbalance problem. This interaction is the foundation of our technique, improving the accuracy and generalization capacities of the detection models and tackling the core difficulty posed by the Medicare Part B dataset.

Algorithm 1 presents the SMOTE-ENN method, specifically tailored for balancing the Medicare Part B Dataset through a combined approach of oversampling minority classes and undersampling majority classes. In its initial phase, the algorithm focuses on oversampling. It randomly selects a minority class instance x_i and determines its k nearest neighbors, thereby creating a subset S_k. A synthetic instance p is then interpolated between x_i and a random member from S_k, which is subsequently labeled as part of the minority class and integrated into the dataset S. This process enhances the minority class's presence, mitigating the imbalance.

The algorithm then transitions to undersampling, aiming to refine the majority class by excising instances likely to introduce classification noise. It selects a random instance x_r from S and identifies its k nearest neighbors. Should x_r predominantly associate with the opposite class, it is pruned,

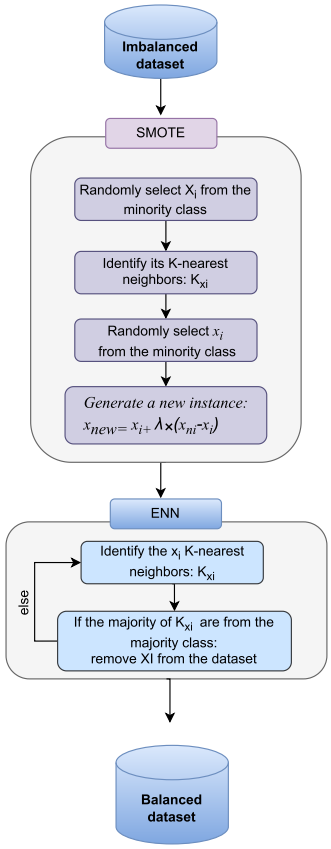

FIGURE 4. SMOTE-ENN process.

reducing the risk of overfitting and bolstering the classifier's generalizability.

The culmination of this two-phase procedure is a balanced dataset S', primed for training resilient machine learning models. By leveraging *SMOTE* for enrichment and *ENN* for purification, the *SMOTE-ENN* algorithm significantly elevates the dataset's utility, thus serving as an essential instrument in optimizing classifier efficacy amidst the complex terrain of healthcare fraud detection.

G. CLASSIFICATION

To classify data as fraud or legitimate, we employ six ML classifiers, namely: Extreme Gradient Boosting (XGBoost), Adaptive Boosting (Adaboost), Light Gradient Boosting Machine (LGBM), Decision Trees (DT), Logistic Regression (LR), and Random Forest classifiers. Ensemble approaches like XGBoost, Adaboost, LightGBM, and RF are widely recognized for their resilience and effectiveness within the field of ML [37].

- Extreme Gradient Boosting (XGBoost): a highly efficient and scalable variant of gradient boosting, recognized for its exceptional performance and speed, making it a fundamental component of our ensemble of classifiers [38].
- Adaptive Boosting (AdaBoost): this method improves the performance of basic models by concentrating on

Algorithm 1 SMOTE-ENN Algorithm for Balancing Medicare Part B Dataset

function SMOTE-ENN($\mathcal{D}_m$)
  Input: Training dataset $\mathcal{D}_m = \{(\mathbf{x}_i^m, y_i^m)\}_{i=1}^{N_m}$
  Output: Balanced dataset $\mathcal{D}'_m$
  Oversampling ▷ Step 1: Oversampling the minority class
  Select a sample $(\mathbf{x}_i^m, y_i^m)$ randomly from minority class instances in $\mathcal{D}_m$
  $S_k \leftarrow$ Find the k nearest minority class neighbors of $\mathbf{x}_i^m$
  $\mathbf{p} \leftarrow$ Generate a synthetic sample by interpolation between $\mathbf{x}_i^m$ and a randomly selected $\mathbf{x}_k$ from S_k
  Assign the minority class label to the new sample $\mathbf{p}$
  Add the new sample $\mathbf{p}$ to the dataset $\mathcal{D}_m$
  Undersampling ▷ Step 2: Undersampling the majority class
  Select a sample $(\mathbf{x}_r^m, y_r^m)$ randomly from $\mathcal{D}_m$
  $S_k \leftarrow$ Find the k nearest neighbors of $\mathbf{x}_r^m$
  if the majority of $\mathbf{x}_r^m$'s neighbors are from the majority class **then**
    Remove $(\mathbf{x}_r^m, y_r^m)$ from $\mathcal{D}_m$
  end if
  $\mathcal{D}'_m \leftarrow$ Balanced dataset
  return $\mathcal{D}'_m$
end function

instances that were incorrectly identified by earlier models [39].

- Light Gradient Boosting Machine (LightGBM): a popular gradient boosting framework known for its efficiency in handling large-scale data while reducing computational resources [40].
- Decision trees (DT): are used for their simplicity and interpretability to classify data by dividing the dataset recursively [41].
- Logistic Regression (LR): Utilizes a logistic function to predict probabilities, offering an accurate model for binary classification tasks like distinguishing between fraudulent and legitimate transactions [42].
- Random Forest (RF): improve predictive accuracy and prevent overfitting by combining predictions from many decision trees, each trained on different subsets of data. This makes them a crucial component of our ensemble strategy [43].

Our choice of these classifiers is based on their combined robustness and effectiveness in the field of machine learning [37]. Ensemble approaches like XGBoost, AdaBoost, LightGBM, and RF excel at combining different models to capture a wider range of patterns and linkages in the data. This characteristic is especially beneficial in healthcare fraud detection, where the intricate and ever-changing fraudulent patterns require advanced, flexible analytical approaches [44]. We utilize the distinct capabilities of each classifier to tackle the difficulties involved in identifying healthcare fraud, guaranteeing the ongoing effectiveness of our model despite changing fraud patterns.

V. EXPERIMENTAL RESULTS

This section provides an evaluation and validation of the performance of the presented models in the detection of healthcare fraud. We utilize a variety of libraries available in the Python programming language, including Pandas, Numpy, and Matplotlib packages from the *sklearn* library. To assess the effectiveness of the models, a series of performance experiments are conducted on the dataset outlined in Section IV. This dataset comprises both valid and fraudulent healthcare claims. The following subsections describe the validation, methodologies, and evaluation metrics employed.

A. VALIDATION

To assess the performance of the proposed models, we employ two common methods: Train_TestSplit and Cross-validation. The Train-Test Split method involves dividing the dataset into two separate subsets: a training set and a testing set. This partitioning enables the evaluation of the model's performance on unseen data. In our work, we adopt an 80:20 split ratio, allocating 80% of the data for training and 20% for testing purposes. On the other hand, k-fold Cross-validation is crucial in the context of healthcare fraud detection, and involves partitioning the data into five distinct subsets. Cross-validation significantly mitigates false positives and negatives, enhancing the model's accuracy in identifying fraud and ensuring a more robust and reliable evaluation.

B. EVALUATION METRICS

Evaluation metrics are important when it comes to assessing the efficacy of ML models in the detection of healthcare fraud. Accuracy, precision, recall, F1 score, and area under the curve (AUC) are frequently used metrics. Specifically, the AUC metric plots the true positive rate against the false positive rate at various threshold settings [45]. Additionally, in the context of the imbalanced dataset, we use the Area Under the Precision-Recall Curve (AUPRC) which offers better insight into the classification performance. It measures the relationship between precision-recall and presents it

in a single value. A higher AUCPR value indicates good performance in correctly identifying positive cases [46]. The rest of the metrics are described as follows:

- **True positive (TP):** a fraud sample is correctly identified as a fraud.
- **True negative (TN):** a non-fraud sample is correctly identified as non-fraud.
- **False positive (FP):** a non-fraud sample is incorrectly identified as fraud.
- **False negative (FN):** a fraud simple is incorrectly identified as non-fraud.
- **Total positive (P):** TP+FN.
- **Total negative (N):** TN+FP
- **Accuracy:** The accuracy metric represents the percentage of occurrences that are correctly classified. It is calculated using the following formula [47]:

$$\frac{TP + TN}{P + N} \tag{2}$$

- **Recall:** Also referred to as sensitivity, the true positive rate is a measure of the proportion of correctly classified instances in the positive class. It is computed using the following formula [48]:

$$\frac{TP}{TP + FN} \tag{3}$$

- **Precision:** it indicates the ratio of the positive samples that are fraud. It is calculated as follows [49]:

$$\frac{TP}{TP + FP} \tag{4}$$

- **F1-score:** it is the weighted average of both Precision and Recall. The F1 score is computed using the following formula [50]:

$$2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

C. RESULTS

The objective of our study is to assess different models and their implications for enhancing fraud detection within the healthcare industry. This section provides a comprehensive analysis and discussion of the outcomes obtained through the implementation of the suggested approach on the Medicare PartB dataset. We apply Logistic Regression LR, DecisionTrees DT, Random Forest RF, XGBoost, Adaboost, and LGBM, as classifiers, with accuracy, F1-score, precision, and recall as evaluation metrics.

1) CLASSIFICATION RESULTS AT BASELINE

Table 3 presents the obtained results of different ML methods to detect healthcare fraud using the Train_Test_split method. While the classifiers have a remarkable accuracy of 0.9999, they exhibit shortcomings in reliably identifying positive instances, as demonstrated by precision, recall, and F1-score values of 0.0000. The AUC values range from a low value of 0.5030 in the case of the DT to a high value of 0.8337 for XGBoost. These values indicate a moderate capability for distinguishing between the two classes. Table 4 displays the obtained results of baseline classification of various ML algorithms using Cross-validation. All the classifiers achieve a perfect accuracy of 0.9999, which signifies good classification of the instances. Nevertheless, the F1-score, precision, and recall metrics for all classifiers continuously exhibit a value of 0.0000. For the AUC values, XGBoost achieves the highest at 0.7444 and RF the lowest at 0.4966. These AUC values still present a significant challenge to effectively discriminating between classes.

TABLE 3. Baseline classification using Train_Test_split.

Classifier	Accuracy	F1-Score	Precision	Recall	AUC
LR	0.9999	0.0000	0.0000	0.0000	0.6200
DT	0.9999	0.0045	0.0062	0.0052	0.5030
RF	0.9999	0.0000	0.0000	0.0000	0.5467
XGBoost	0.9999	0.0000	0.0000	0.0000	0.8337
Adaboost	0.9999	0.0000	0.0000	0.0000	0.8073
LGBM	0.9998	0.0000	0.0000	0.0000	0.7986

TABLE 4. Baseline classification using cross-validation.

Classifier	Accuracy	F1-Score	Precision	Recall	AUC
LR	0.9999	0.0000	0.0000	0.0000	0.6044
DT	0.9998	0.0000	0.0000	0.0000	0.4999
RF	0.9999	0.0000	0.0000	0.0000	0.4966
XGBoost	0.9990	0.0000	0.0000	0.0000	0.7444
Adaboost	0.9999	0.0000	0.0000	0.0000	0.7374
LGBM	0.9998	0.0000	0.0000	0.0000	0.7039

TABLE 5. Classification results using SMOTE-ENN and Train_Test_split.

Classifier	Accuracy	F1-Score	Precision	Recall	AUC
LR	0.65	0.68	0.57	0.83	0.67
DT	0.99	0.99	0.99	0.99	0.99
RF	0.90	0.90	0.82	0.99	0.94
XGBoost	0.95	0.95	0.94	0.96	0.95
Adaboost	0.64	0.69	0.56	0.89	0.67
LGBM	0.90	0.89	0.88	0.90	0.90

2) CLASSIFICATION RESULTS USING SMOTE-ENN

Table 5 presents the obtained classification results using SMOTE-ENN and train-test-split methods. We can observe that the DT classifier exhibits the highest performance across all metrics, with accuracy, F1-score, precision, recall, and AUC, each at 0.99. XGBoost also presents good results with an accuracy, F1-score, and AUC of 0.95, a precision of 0.94, and a recall of 0.96. RF and LBGM classifiers achieve a similar accuracy of 0.90; RF achieves an F1-score of 0.90, a precision of 0.82, a 0.94 AUC, and a high recall of 0.99. Whereas LGBM obtains an F1-score of 0.89, precision of 0.88, recall, and AUC of 0.90. Conversely, LR and Adaboost demonstrate relatively poor results. LR obtains an accuracy of 0.65, an F1-score of 0.68, a precision of 0.57, a 0.67 AUC, and a high recall of 0.83. For Adaboost, the results show a low accuracy of 0.64, an F1-score of 0.69, a precision of 0.56, a 0.67 AUC, and a notably high recall of 0.89. Table 8 displays the classification results using SMOTE-ENN and cross-validation. DT outperforms the other classifiers with

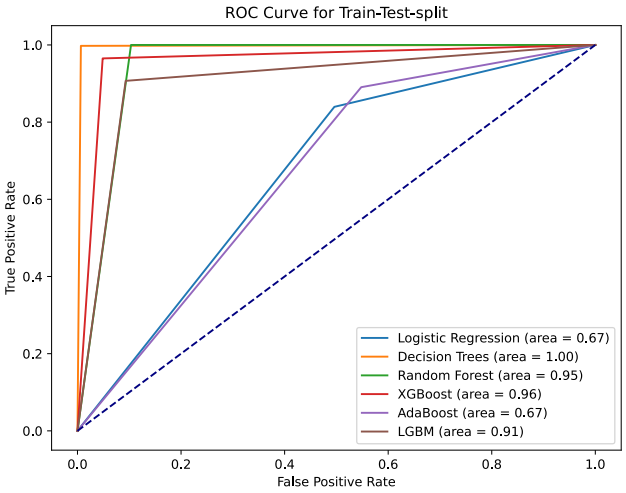

FIGURE 5. ROC curve for each model in train-test-split.

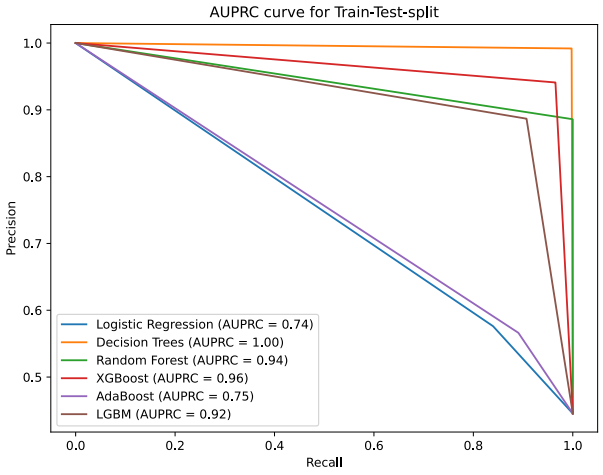

FIGURE 6. AUPRC curve for each model in train-test-split.

TABLE 6. Classification results using SMOTE-ENN and cross-validation.

Classifier	Accuracy	F1-Score	Precision	Recall	AUC
LR	0.65	0.65	0.69	0.67	0.73
DT	1.00	1.00	0.99	1.00	0.95
RF	0.95	0.95	0.95	0.95	0.99
XGBoost	0.96	0.96	0.96	0.96	0.99
Adaboost	0.65	0.64	0.70	0.67	0.68
LGBM	0.91	0.91	0.90	0.91	0.97

accuracy, F1-score, and recall of 1.00, precision of 0.99, and 0.95 AUC. Following this, XGBoost attains remarkable results, with an accuracy, F1-score, precision, recall of 0.96, and a good AUC value of 0.99. RF presents good results, with 0.95 for all metrics and 0.99 for AUC. LGBM also presents good results with 0.91 accuracy, F1-score, and recall. The model also achieves a precision of 0.90 and an AUC of 0.97. Simultaneously, LR and Adaboost classifiers attain closely similar outcomes, with an accuracy and F1-score of 0.65, precision of 0.69, recall of 0.67, and a 0.73 AUC for LR. For AdaBoost, the metrics indicate an accuracy of 0.65, an F1-score of 0.64, a precision of 0.70, a recall of 0.67, and an AUC score of 0.68.

D. DISCUSSION

This paper introduces a new ML approach, employing the hybrid resampling method *SMOTE-ENN* to tackle the imbalanced data problem within the Medicare dataset. Additionally, it investigates the unique treatment of categorical features alongside numerical data to enhance the efficacy of the fraud detection process. This approach demonstrated efficacy compared to traditional techniques such as ROS, RUS, and the basic SMOTE method, particularly in its proficiency in handling imbalanced data while reducing noisy data. The experiments show significant performance variations among different ML models.

1) DISCUSSION OF THE BASELINE CLASSIFICATION RESULTS

The initial baseline results from both Train_test_split and cross-validation methods exhibit high accuracy (0.9999) across all classifiers. Nevertheless, these high results of accuracy are misleading, especially when facing a highly imbalanced dataset. When the majority class dominates the minority class, the models always predict the majority class without actually learning to identify the characteristics of the minority class, in this case, the fraudulent instances. This is highlighted by the low results of precision, recall, and F1-score among all classifiers, underscoring their ineffectiveness in accurately identifying fraud instances. Moreover, while AUC scores show some improvement, reaching the highest score of 0.8337 by XGBoost, these values are still not optimal for reliable fraud detection.

In the Appendix A, we present the obtained results of baseline classification utilizing train-test splits with two ratios, specifically 25:75 and 30:70. The baseline classifications for the 25:75 and 30:70 ratios showed high accuracy with LR, DT, RF, XGBoost, Adaboost, and LGBM classifiers. However, they had notably low F1-Scores, Precision, Recall, and AUC values, suggesting a limited ability to predict the minority class effectively. These findings indicate that a severe class imbalance can significantly impact the results of the classification task.

2) DISCUSSION OF THE CLASSIFICATION RESULTS USING SMOTE-ENN

The SMOTE-ENN technique, which generates new minority class instances and removes overlapping samples from the dataset, has varying effects on different classifiers. It notably improves tree-based and ensemble methods but has limited influence on LR and Adaboost. For instance, DT exhibits perfect results with 0.99 across all metrics with Train_test_split. Similarly, for cross-validation, the model attains perfect results of 1.00 with accuracy, F1-score, and recall. These strong results affirm the effectiveness of the proposed approach to boost the classifier's performance on imbalanced data sets. The difference in the algorithms' handling of imbalanced data and synthesized instances is due to LR's linearity and Adaboost's sensitivity to noise, which prevent

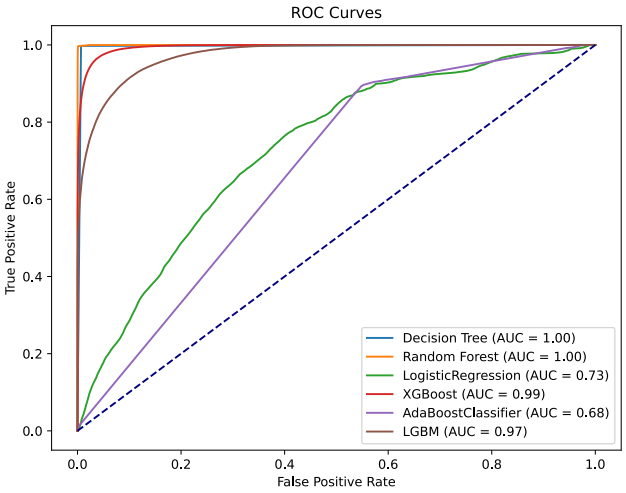

FIGURE 7. ROC curve for each model in cross-validation.

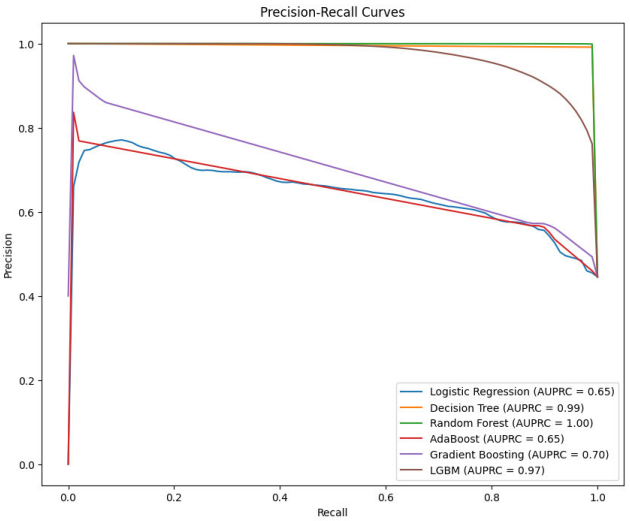

FIGURE 8. AUPRC curve for each model in cross-validation.

them from fully utilizing over-sampling benefits. Our strategy focuses on enhancing fraud detection by adjusting class distribution and strengthening the models' capacity to learn from the minority class. Creating new fraudulent transactions and removing noisy data helps improve generalization from training to unseen data, enhancing the effectiveness of fraud detection across different classification methods.

3) DISCUSSION OF AUPRC CURVE

Another important point in the imbalanced data is the analysis of the ROC-AUC and AUPRC curves, which plays a crucial role in understanding the performance of ML models in the context of imbalanced datasets for healthcare fraud detection. The ROC-AUC curve in figures 6 and 8 reinforces our initial findings, notably highlighting the perfect results of DT. This consistency between the ROC curve results and our initial findings presents a comprehensive validation of the model's performance, particularly in the context of imbalanced datasets in healthcare fraud detection. Moreover, the AUPRC curves in 6 and 8, which are relevant in the case of imbalanced data, confirm the obtained results. These findings from the ROC-AUC and AUPRC curves are important to understanding our models since they give a clear overview of each model's strengths and weaknesses.

4) COMPARAISON WITH STATE-OF-THE-ART

Our research significantly contributes to the healthcare fraud field by addressing the imbalanced data problem. By utilizing the SMOTE-ENN technique, along with the generation of the categorical feature, we have enhanced the ability of ML models to accurately identify instances of fraud. This method surpasses traditional techniques like ROS, RUS, and basic SMOTE, as well as other studies that focus on different sampling techniques or embeddings. Our research, therefore, contributes to the ongoing efforts to develop effective systems to detect fraud in healthcare.

Tables 10, 11, and 12 in Appendix B B present the obtained results of different classifiers (LR, DT, RF, XGBoost, Adaboost, and LGBM) using three distinct data sampling methods: RUS, ROS, and SMOTE combined with Train_Test_split for the classification task. For instance, when using RUS, performance measures show a respectable level of accuracy, with the LGBM classifier slightly outperforming in accuracy (0.74), recall (0.77), and AUC (0.75). Nevertheless, all classifiers face challenges in achieving precision and F1-score, with both metrics registering zero. RUS may enhance model sensitivity but significantly reduce precision, resulting in a high rate of false positives. Conversely, ROS significantly improves accuracy for RF and XGBoost models but fails to enhance precision or recall for minority class predictions, revealing a crucial limitation in detecting the minority class. SMOTE provides the highest accuracy, especially for DT and RF, yet does not address the problem of near-zero precision and recall.

SMOTE-ENN excels at handling imbalanced datasets by effectively balancing and removing noisy data, surpassing approaches such as RUS, ROS, and SMOTE. This method effectively addresses the limitations of singular balancing methods, providing a sophisticated approach to improve the classifier's performance in the presence of class imbalances. Table 7 compares the best outcomes achieved by each method, specifically RUS, ROS, SMOTE, and our proposed methodology. Our methodology outperformed standard methods in all evaluated measures, demonstrating its effectiveness.

Figure 9 illustrates a comparison between RUS, ROS, SMOTE, and our proposed approach, highlighting the superior performance of our technique in terms of accuracy, F1-score, precision, recall, and AUC, surpassing traditional methods significantly.

5) LIMITATIONS

Despite this, our research methodology has certain limitations that need to be addressed. While this approach demonstrates

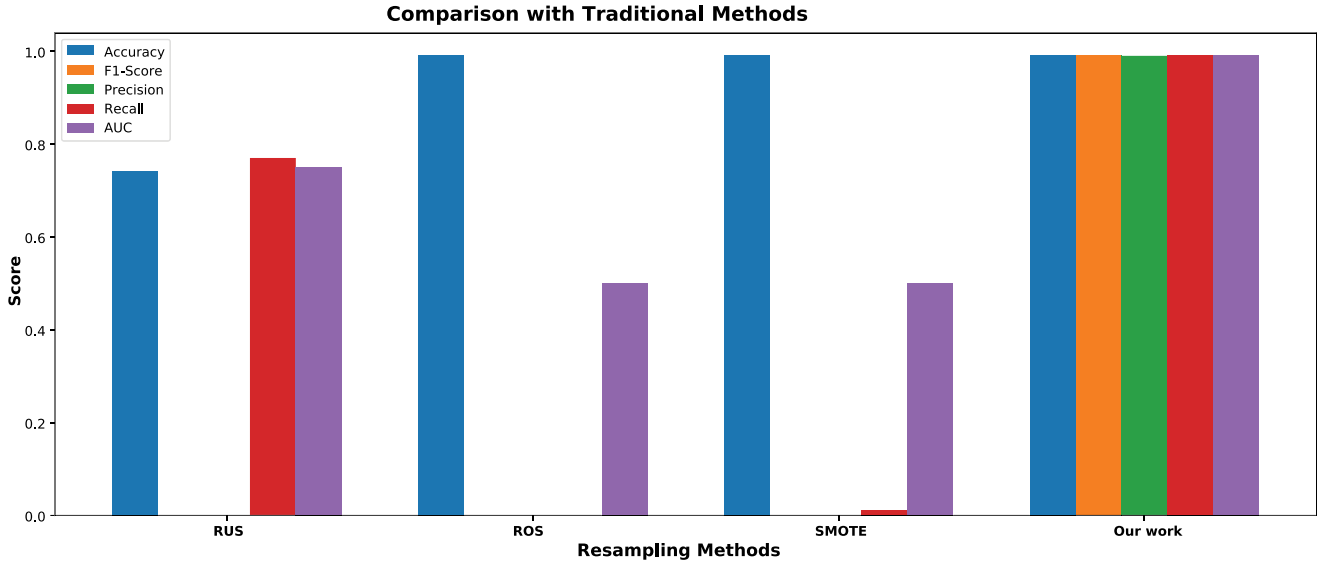

FIGURE 9. Comparison with traditional methods.

TABLE 7. Comparison with traditional methods.

Classifier	Accuracy	F1-Score	Precision	Recall	AUC
RUS	0.74	0.00	0.00	0.77	0.75
ROS	0.99	0.00	0.00	0.00	0.50
SMOTE	0.99	0.00	0.00	0.01	0.50
Our work	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**

remarkable results, it was not uniformly observed across all ML models. In addition, the study relies on the Medicare PartB dataset, which might limit the generalizability of our findings to other datasets. Future work could explore several promising directions. One direction is the exploration of a new dataset, such as Medicare PartD, that encompasses prescription drug benefits. This presents a different set of challenges and patterns of fraud compared to other parts of Medicare. Moreover, advanced deep learning models like Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) show potential for addressing imbalanced datasets [15]. These models are adept at capturing intricate patterns in extensive datasets, making them especially well-suited for applications where the minority class is vital, such as fraud detection. LSTM networks, a type of RNN, are highly effective at assessing sequential data, such as time-series medical billing information or patient treatment sequences, to identify anomalies or fraudulent trends [51]. Similarly, Convolutional Neural Networks (CNNs) excel at handling class imbalance by leveraging their feature extraction capabilities, particularly in high-dimensional data spaces like images and signal processing. This renders them appropriate for complex tasks that challenge traditional techniques, such as medical imaging for diagnosing rare conditions among several normal cases. Their hierarchical learning method efficiently reveals important patterns in imbalanced datasets, demonstrating its adaptability in various fields, such as natural language processing and healthcare diagnostics [52]. By exploring these avenues, future research can build upon our findings, potentially leading to more robust and comprehensive fraud detection systems in healthcare.

VI. CONCLUSION

This study emphasizes the need to address imbalanced data in healthcare fraud detection by introducing a novel ML framework based on the SMOTE-ENN hybrid resampling method. This method effectively balances datasets by creating synthetic samples while eliminating noisy data, thereby enhancing the model's accuracy. Another aspect of our study is the application of the AUC and AUPRC as evaluation metrics. These metrics facilitated a thorough analysis of the models' performance, with the AUPRC proving to be especially critical in the context of imbalanced datasets. Thus, this approach serves as a basis for new researchers to apply new approaches to detect healthcare fraud. Future research directions include evaluating SMOTE-ENN's performance in diverse healthcare fraud scenarios and combining it with innovative AI technologies such as deep learning (DL) to enhance the effectiveness of fraud detection methods.

APPENDIX A
BASELINE CLASSIFICATION WITH DIFFERENT TRAIN_TEST_SPLIT RATIOS

The appendix provides two tables demonstrating the performance of six different classifiers (LR, DT, RF, XGBoost, Adaboost, and LGBM) across two training-to-testing ratios (25:75 and 30:70) on the Medicare PartB dataset. Both tables 8 and 9 present a comparative comparison across various metrics: Accuracy, F1-Score, Precision, Recall, and AUC.

TABLE 8. Baseline classification using 25:75 ratio Train_Test_split.

Classifier	Accuracy	F1-Score	Precision	Recall	AUC
LR	0.9999	0.0000	0.0000	0.0000	0.4999
DT	0.9999	0.0045	0.0062	0.0052	0.4999
RF	0.9999	0.0000	0.0000	0.0000	0.4999
XGBoost	0.9999	0.0000	0.0000	0.0000	0.5
Adaboost	0.9999	0.0000	0.0000	0.0000	0.4999
LGBM	0.9998	0.0000	0.0000	0.0000	0.4999

TABLE 9. Baseline classification using 30:70 ratio Train_Test_split.

Classifier	Accuracy	F1-Score	Precision	Recall	AUC
LR	0.9999	0.0000	0.0000	0.0000	0.4999
DT	0.9999	0.0003	0.0003	0.0004	0.5
RF	0.9999	0.0000	0.0000	0.0000	0.4999
XGBoost	0.9999	0.0000	0.0000	0.0000	0.5
Adaboost	0.9999	0.0000	0.0000	0.0000	0.4999
LGBM	0.9998	0.0003	0.0003	0.0004	0.502

TABLE 10. Classification results using RUS and Train_Test_split.

Classifier	Accuracy	F1-Score	Precision	Recall	AUC
LR	0.45	0.00	0.00	0.85	0.65
DT	0.69	0.00	0.00	0.66	0.67
RF	0.72	0.00	0.00	0.77	0.75
XGBoost	0.73	0.00	0.00	0.75	0.74
Adaboost	0.68	0.00	0.00	0.76	0.72
LGBM	**0.74**	0.00	0.00	**0.77**	**0.75**

TABLE 11. Classification results using ROS and Train_Test_split.

Classifier	Accuracy	F1-Score	Precision	Recall	AUC
LR	0.52	0.00	0.00	0.83	0.67
DT	0.52	0.00	0.00	0.66	0.67
RF	**0.99**	0.00	0.00	0.00	0.5
XGBoost	0.96	0.00	0.00	0.21	0.58
Adaboost	0.71	0.00	0.00	0.75	0.72
LGBM	0.74	0.00	0.00	0.77	**0.75**

TABLE 12. Classification results using SMOTE and Train_Test_split.

Classifier	Accuracy	F1-Score	Precision	Recall	AUC
LR	0.52	0.00	0.00	0.74	0.63
DT	**0.99**	0.00	0.00	0.01	0.50
RF	**0.99**	0.00	0.00	0.00	0.49
XGBoost	0.95	0.00	0.00	0.21	0.58
Adaboost	0.67	0.00	0.00	0.70	**0.69**
LGBM	0.88	0.00	0.00	0.37	0.62

APPENDIX B
CLASSIFICATION RESULTS USING STATE-OF-THE-ART RESAMPLING METHODS

This appendix presents the results of the classification utilizing six algorithms as well as traditional resampling methods, including ROS, RUS, and SMOTE. Table 10 presents the classification results using the RUS method; Table 11 shows the experiment results using ROS; and Table 12 outlines the obtained results using the SMOTE method.

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: http://ieeexplore.ieee.org/document/8260744/
[6] V. Nalluri, J.-R. Chang, L.-S. Chen, and J.-C. Chen, "Building prediction models and discovering important factors of health insurance fraud using machine learning methods," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 7, pp. 9607–9619, Jul. 2023.
[7] P. Dua and S. Bais, "Supervised learning methods for fraud detection in healthcare insurance," in *Machine Learning in Healthcare Informatics* (Intelligent Systems Reference Library), vol. 56, S. Dua, U. Acharya, and P. Dua, Eds. Berlin, Germany: Springer, 2014, doi: 10.1007/978-3-642-40017-9_12.
[8] R. Bauder, R. da Rosa, and T. Khoshgoftaar, "Identifying medicare provider fraud with unsupervised machine learning," in *Proc. IEEE Int. Conf. Inf. Reuse Integr. (IRI)*, Jul. 2018, pp. 285–292.
[9] Centers for Medicare and Medicaid Services. (2017). *Research, Statistics, Data, and Systems*. [Online]. Available: https://www.cms.gov/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.
[11] N. Agrawal and S. Panigrahi, "A comparative analysis of fraud detection in healthcare using data balancing & machine learning techniques," in *Proc. Int. Conf. Commun., Circuits, Syst. (IC3S)*, May 2023, pp. 1–4.
[12] M. Herland, R. A. Bauder, and T. M. Khoshgoftaar, "The effects of class rarity on the evaluation of supervised healthcare fraud detection models," *J. Big Data*, vol. 6, no. 1, pp. 1–33, Dec. 2019.
[13] J. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, "The effects of random undersampling for big data medicare fraud detection," in *Proc. IEEE Int. Conf. Service-Oriented Syst. Eng. (SOSE)*, Aug. 2022, pp. 141–146.
[14] A. Mehbodniya, I. Alam, S. Pande, R. Neware, K. P. Rane, M. Shabaz, and M. V. Madhavan, "Financial fraud detection in healthcare using machine learning and deep learning techniques," *Secur. Commun. Netw.*, vol. 2021, pp. 1–8, Sep. 2021.
[15] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017.
[16] J. Hancock and T. M. Khoshgoftaar, "Optimizing ensemble trees for big data healthcare fraud detection," in *Proc. IEEE 23rd Int. Conf. Inf. Reuse Integr. Data Sci. (IRI)*, Aug. 2022, pp. 243–249.
[17] N. Kumaraswamy, M. K. Markey, J. C. Barner, and K. Rascati, "Feature engineering to detect fraud using healthcare claims data," *Expert Syst. Appl.*, vol. 210, Dec. 2022, Art. no. 118433.
[18] N. Kumaraswamy, T. Ekin, C. Park, M. K. Markey, J. C. Barner, and K. Rascati, "Using a Bayesian belief network to detect healthcare fraud," *Expert Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 122241.
[19] J. M. Johnson and T. M. Khoshgoftaar, "Data-centric AI for healthcare fraud detection," *Social Netw. Comput. Sci.*, vol. 4, no. 4, p. 389, May 2023.
[20] R. A. Bauder and T. M. Khoshgoftaar, "The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data," *Health Inf. Sci. Syst.*, vol. 6, no. 1, pp. 1–14, Dec. 2018.
[21] R. A. Bauder, T. M. Khoshgoftaar, and T. Hasanin, "Data sampling approaches with severely imbalanced big data for medicare fraud detection," in *Proc. IEEE 30th Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2018, pp. 137–142.
[22] J. M. Johnson and T. M. Khoshgoftaar, "Hcpcs2Vec: Healthcare procedure embeddings for medicare fraud prediction," in *Proc. IEEE 6th Int. Conf. Collaboration Internet Comput. (CIC)*, Dec. 2020, pp. 145–152.
[23] J. M. Johnson and T. M. Khoshgoftaar, "Medical provider embeddings for healthcare fraud detection," *Social Netw. Comput. Sci.*, vol. 2, no. 4, p. 276, Jul. 2021. [Online]. Available: https://link.springer.com/10.1007/s42979-021-00656-y

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

[25] J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, "Evaluating classifier performance with highly imbalanced big data," *J. Big Data*, vol. 10, no. 1, p. 42, Apr. 2023.

[26] J. T. Hancock and T. M. Khoshgoftaar, "Exploring maximum tree depth and random undersampling in ensemble trees to optimize the classification of imbalanced big data," *Social Netw. Comput. Sci.*, vol. 4, no. 5, p. 462, Jun. 2023.

[27] CMS. *Medicare Physician & Other Practitioners–By Provider—Centers for Medicare & Medicaid Services Data*. Accessed: Aug. 3, 2023. [Online]. Available: https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners/medicare-physician-other-practitioners-by-provider

[28] OIG. *LEIE Downloadable Databases | Office of Inspector General | US Department of Health and Human Services*. Accessed: Aug. 4, 2023. [Online]. Available: https://oig.hhs.gov/exclusions/exclusions_list.asp

[29] CMS. *Medicare Physician & Other Practitioners Methodology—Centers for Medicare & Medicaid Services Data*. Accessed: Aug. 24, 2023. [Online]. Available: https://data.cms.gov/resources/medicare-physician-other-practitioners-methodology

[30] R. Bauder and T. Khoshgoftaar, "Medicare fraud detection using random forest with class imbalanced big data," in *Proc. IEEE Int. Conf. Inf. Reuse Integr. (IRI)*, Jul. 2018, pp. 80–87.

[31] J. Hancock and T. M. Khoshgoftaar, "Medicare fraud detection using CatBoost," in *Proc. IEEE 21st Int. Conf. Inf. Reuse Integr. Data Sci. (IRI)*, Aug. 2020, pp. 97–103.

[32] M. Herland, T. M. Khoshgoftaar, and R. A. Bauder, "Big data fraud detection using multiple medicare data sources," *J. Big Data*, vol. 5, no. 1, pp. 1–21, Dec. 2018.

[33] M. Rashid, J. Kamruzzaman, T. Imam, S. Wibowo, and S. Gordon, "A tree-based stacking ensemble technique with feature selection for network intrusion detection," *Appl. Intell.*, vol. 52, no. 9, pp. 9768–9781, 2022.

[34] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newslett.*, vol. 6, no. 1, pp. 20–29, Jun. 2004.

[35] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[36] I. D. Mienye and Y. Sun, "A deep learning ensemble with data resampling for credit card fraud detection," *IEEE Access*, vol. 11, pp. 30628–30638, 2023.

[37] J. Ye, J.-H. Chow, J. Chen, and Z. Zheng, "Stochastic gradient boosted distributed decision trees," in *Proc. 18th ACM Conf. Inf. Knowl. Manage.*, Nov. 2009, pp. 2061–2064.

[38] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, and T. Zhou, "XGBoost: Extreme gradient boosting," *R Package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[39] Y. Cao, Q.-G. Miao, J.-C. Liu, and L. Gao, "Advance and prospects of AdaBoost algorithm," *Acta Automatica Sinica*, vol. 39, no. 6, pp. 745–758, Mar. 2014.

[40] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–9.

[41] S. B. Kotsiantis, "Decision trees: A recent overview," *Artif. Intell. Rev.*, vol. 39, no. 4, pp. 261–283, Apr. 2013.

[42] S. Lemeshow, R. X. Sturdivant, and D. W. Hosmer Jr., *Applied Logistic Regression*. Hoboken, NJ, USA: Wiley, 2013.

[43] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, Jun. 2016.

[44] J. T. Hancock and T. M. Khoshgoftaar, "Gradient boosted decision tree algorithms for medicare fraud detection," *Social Netw. Comput. Sci.*, vol. 2, no. 4, p. 268, Jul. 2021.

[45] S. Wu and P. Flach, "A scored AUC metric for classifier evaluation and selection," in *Proc. 2nd Workshop ROC Anal. ML*, Bonn, Germany, 2005, pp. 1–3.

[46] K. Boyd, K. H. Eng, and C. D. Page, "Area under the precision-recall curve: Point estimates and confidence intervals," in *Machine Learning and Knowledge Discovery in Databases* (Lecture Notes in Computer Science), vol. 8190, H. Blockeel, K. Kersting, S. Nijssen, and F. Železný, Eds. Berlin, Germany: Springer, 2013, doi: 10.1007/978-3-642-40994-3_29.

[47] P. Y. Prasad, A. S. Chowdary, C. Bavitha, E. Mounisha, and C. Reethika, "A comparison study of fraud detection in usage of credit cards using machine learning," in *Proc. 7th Int. Conf. Trends Electron. Informat. (ICOEI)*, Apr. 2023, pp. 1204–1209.

[48] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation measures for models assessment over imbalanced data sets," *J. Inf. Eng. Appl.*, vol. 3, no. 10, pp. 27–38, 2013.

[49] P. Gupta, A. Varshney, M. R. Khan, R. Ahmed, M. Shuaib, and S. Alam, "Unbalanced credit card fraud detection data: A machine learning-oriented comparative study of balancing techniques," *Proc. Comput. Sci.*, vol. 218, pp. 2575–2584, 2023.

[50] B. Guelib, K. Zarour, H. Hermessi, B. Rayene, and K. Nawres, "Same-subject-modalities-interactions: A novel framework for MRI and PET multi-modality fusion for Alzheimer's disease classification," *IEEE Access*, vol. 11, pp. 48715–48738, 2023.

[51] R. Ghosh, S. Phadikar, N. Deb, N. Sinha, P. Das, and E. Ghaderpour, "Automatic eyeblink and muscular artifact detection and removal from EEG signals using k-nearest neighbor classifier and long short-term memory networks," *IEEE Sensors J.*, vol. 23, no. 5, pp. 5422–5436, Mar. 2023.

[52] D. Dablain et al., "Understanding CNN fragility when learning with imbalanced data," *Mach. Learn.*, 2023, doi: 10.1007/s10994-023-06326-9.

RAYENE BOUNAB received the degree in computer science from Université Abdelhamid Mehri Constantine 2, in 2016, and the master's degree from the Faculty of Nouvelle Téchnologies d'Informations et Communication (NTIC), in 2019. She is currently pursuing the Ph.D. degree with the Department of TLSI, LIRE Laboratory, Université Abdelhamid Mehri Constantine 2. She is also working on machine learning for healthcare fraud detection.

KARIM ZAROUR received the Ph.D. degree from the Mentouri University of Constantine and the Habilitation degree from Université Abdelhamid Mehri Constantine 2, Algeria. He is currently a Professor in computer science with Université Abdelhamid Mehri Constantine 2. He supervises many Ph.D. and master's students. He has published many articles in international journals and conferences. His current research interests include health informatics, IA, privacy and security in healthcare, and multi-agent systems and cloud.

BOUCHRA GUELIB received the B.S. degree from the Faculty of Nouvelle Téchnologies d'Informations et Communication, Université Abdelhamid Mehri Constantine 2, in 2015, and the M.S. degree in information systems from Université Abdelhamid Mehri Constantine 2, in 2018, where she is currently pursuing the Ph.D. degree with the Department of TLSI, LIRE Laboratory. She is also working on multimodal fusion using machine learning. Her research interests include medical image processing, multimodal fusion, and the bioinformatic field.

NAWRES KHLIFA received the Ph.D. degree from the National School of Engineers of Tunis (ENIT). She was an Engineer. She is currently a Professor with the Higher Institute of Medical Technologies of Tunis, Tunis El Manar University. She coordinates the TIMEd Team: Medical Image Processing, BTM Laboratory. Her research interests include artificial intelligence and CAD design in medical imaging, emotion recognition, and gaze tracking.