IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 5, 2021, accepted July 17, 2021, date of publication August 3, 2021, date of current version August 11, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3102399

A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data

MATLOOB KHUSHI[1,2], KAMRAN SHAUKAT[3,4], TALHA MAHBOOB ALAM[5], IBRAHIM A. HAMEED[6], SHAHADAT UDDIN[7], SUHUAI LUO[3], XIAOYAN YANG[1], (Member, IEEE), AND MARANATHA CONSUELO REYES[1]

[1]School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia
[2]School of EAST, University of Suffolk, Ipswitch IP4 1QJ, U.K.
[3]School of Information and Physical Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia
[4]Department of Data Science, University of the Punjab, Lahore 54590, Pakistan
[5]Department of Computer Science and Information Technology, Virtual University of Pakistan, Lahore 54000, Pakistan
[6]Department of ICT and Natural Sciences, Norwegian University of Science and Technology, 7491 Trondheim, Norway
[7]School of Project Management, The University of Sydney, Sydney, NSW 2006, Australia

Corresponding authors: Matloob Khushi (matloobkhushi@gmail.com), Kamran Shaukat (kamran.shaukat@uon.edu.au), and Ibrahim A. Hameed (ibib@ntnu.no)

ABSTRACT Medical datasets are usually imbalanced, where negative cases severely outnumber positive cases. Therefore, it is essential to deal with this data skew problem when training machine learning algorithms. This study uses two representative lung cancer datasets, PLCO and NLST, with imbalance ratios (the proportion of samples in the majority class to those in the minority class) of 24.7 and 25.0, respectively, to predict lung cancer incidence. This research uses the performance of 23 class imbalance methods (resampling and hybrid systems) with three classical classifiers (logistic regression, random forest, and LinearSVC) to identify the best imbalance techniques suitable for medical datasets. Resampling includes ten under-sampling methods (RUS, etc.), seven over-sampling methods (SMOTE, etc.), and two integrated sampling methods (SMOTEENN, SMOTE-Tomek). Hybrid systems include (Balanced Bagging, etc.). The results show that class imbalance learning can improve the classification ability of the model. Compared with other imbalanced techniques, under-sampling techniques have the highest standard deviation (SD), and over-sampling techniques have the lowest SD. Over-sampling is a stable method, and the AUC in the model is generally higher than in other ways. Using ROS, the random forest performs the best predictive ability and is more suitable for the lung cancer datasets used in this study. The code is available at https://mkhushi.github.io/

INDEX TERMS Class imbalance, data resampling, healthcare, lung cancer, machine learning.

I. INTRODUCTION

In a class-imbalanced dataset, one of its classes has a significantly lower number of samples than the other [1]. There are challenges inherent in learning such class imbalanced data. The skewed distribution of the training examples makes standard learning classifiers biased, favouring the majority class and cannot detect rare instances [2], [3]. Rare minority samples may be treated as noise, and noise may be incorrectly identified as minority samples [4], [5]. In the medical field, this type of imbalance problem often exists. The number of normal samples in the dataset is often more than that of abnormal samples, and the gap between the two is relatively large [6]. Researchers have developed various class imbalance methods and performance evaluation metrics to address these challenges, briefly discussed in Section II-A and Section II-B, respectively. The most commonly used abbreviations are presented in Table 1. To investigate class imbalance methods, we implemented them on two real-world class imbalanced datasets: (i) the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial dataset and (ii) the National Lung Screening Trial (NLST) dataset. PLCO and NLST are high-profile datasets in the field of lung cancer, and many researchers have done some research on them [7], [8]. Both datasets contain anonymised clinical information from trial participants, including whether they have confirmed lung cancer or not. In these lung cancer datasets, the ratio of most samples (normal people) to a few

The associate editor coordinating the review of this manuscript and approving it for publication was Wentao Fan.

TABLE 1. List of acronyms.

ADASYN	Adaptive Synthetic
AUC	Area under curve
BB	Balanced Bagging
BRF	Balanced Random Forest
CC	Cluster Centroids
CNN	Condensed Nearest Neighbour
ENN	Edited Nearest Neighbors
FN	False Negative
FP	False Positive
IHT	Instance Hardness Threshold
k-NN	*k*-Nearest Neighbors
NCR	Neighbourhood Cleaning Rule
NLST	National Lung Screening Trial
NM	Near Miss
OSS	One-Sided Selection
PLCO	the Prostate, Lung, Colorectal, and Ovarian
ROS	Random Over-sampling
RUS	Random Under-Sampling
SD	Standard Deviation
SMOTE	Synthetic Minority Over-sampling Technique
SMOTE-NC	Synthetic Minority Over-sampling Technique - Nominal Continuous
ROC	Receiver Operating Characteristic
RUSBoost	Random Under-Sampling Boost
SVM	Support Vector Machine
TL	Tomek Links
TN	True Negative
TP	True Positive

samples (lung cancer patients) is around 25. Therefore, they all belong to the class imbalance dataset, which can explore the class imbalance methods.

II. CLASS SKEWNESS IN DATA

Class skewness is a well-known problem in machine learning [9]. Suppose the distribution of the class in the data is imbalanced. In that case, the machine learning model will tilt towards the samples in the majority class and cannot give enough attention to the samples in the minority class. It will cause the model's output to be biased towards the majority class [10], [11]. The accuracy of the classifier is unreliable due to the lack of consideration of minority classes. In the current field of machine learning, the class skewness in data has caused many scholars to pay attention to class-imbalanced learning [12], [13].

A. TYPES OF IMBALANCED METHODS

In the Biomedical Sciences, class imbalance methods have already been used in many applications, such as gene expression [14], medical diagnosis [15] and medical side effects [16]. Class imbalanced data methods can be classified into three categories: (i) data-level methods, (ii) algorithm-level methods and (iii) hybrid methods [17].

1) DATA-LEVEL METHODS

Data-level methods involve procedures applied in the training data to make the class distribution more balanced by reducing the number of samples in more classes or increasing the number of samples in minority classes [18]. At present, the data-level method is mainly in the data pre-processing stage, using resampling to redistribute the training data of different classes in the data space [19], [20]. This kind of method can change the dataset structure as much as possible to balance the imbalanced class. Some studies have shown that the resampling method can improve the model's ability to a certain extent by resampling the data samples to adjust the analog distribution of the samples [21], [22]. In the data-level method, resampling and the work of the classifier do not affect each other, which is also one of its advantages [23]. Resampling procedures can be further organised into (i) under-sampling, (ii) over-sampling and (iii) hybrid methods [24]. In the following, we briefly describe these methods.

In Under-sampling methods, samples from the majority class are discarded until the number of samples in each class are nearly equal while preserving valuable information for learning [25], [26]. However, it is inevitable that when under-sampling the dataset, some samples that are meaningful to the training model may be ignored [27], [28]. After all, different under-sampling methods have different filtering principles. Under-sampling methods include:

1) *Random Under-Sampling (RUS):* RUS is the earliest under-sampling technique developed; it discards random samples from the majority class 29].
2) *All k-Nearest Neighbors (All k-NN):* For all values from 1 until the given value of k, this method performs k-NN to each sample. If the majority of its neighbours classify an instance incorrectly, that instance is discarded [30].
3) *Cluster Centroids:* This method performs k-means and replaces the majority class samples with their respective cluster centroids to reduce the number of samples [31].
4) *Edited Nearest Neighbors (ENN):* Each instance is tested using k-NN with the rest of the samples in this method. Those incorrectly classified will be discarded, and the remaining samples will form the edited dataset [32].
5) *Instance Hardness Threshold (IHT):* This under-sampling method first trains a classifier to determine hard instances or those with a high probability of being misclassified, then removes them [33].
6) *Near Miss:* This technique selects majority samples close to some minority samples; that is, their average distances to the three closest minority samples are smallest [34].
7) *Neighbourhood Cleaning Rule (NCR):* This method considers three nearest neighbours of each instance in the dataset. If a sample belongs to the majority class and is misclassified by its three nearest neighbours, it is removed from the dataset. Also, if a sample belongs to the minority class sample and is misclassified by its three nearest neighbours, then the majority class samples among its neighbours are removed [35].
8) *One-Sided Selection (OSS):* First, minority class samples and misclassified majority samples are selected by 1-NN. Then a majority of class samples in the Tomek Links are removed [36].

9) *Repeated ENN:* This method performs ENN repeatedly until the edited training set becomes unaffected by further elimination [37].
10) *Tomek Links (TL):* Two instances a and b are Tomek Links if they belong to different classes and are one another's nearest neighbour. Thus, Tomek Links are boundary or noisy instances, and the sample from the majority class is removed [38].
11) *Condensed Nearest Neighbour (CNN):* Use the nearest neighbour algorithm to iterate, and use under-sampling to put the majority class sample and all the minority class samples together into a set C. The remaining part uses 1-NN to judge whether it can be classified correctly, and the wrongly classified samples are put into set C. Repeat the above process to determine whether the majority class of samples can be retained [39].

In over-sampling methods, new samples are created based on samples from the minority class to reach a more balanced class distribution of samples while strengthening class boundaries [40], [41]. However, over-sampling may lead to overfitting because it duplicates or synthesises a minority of samples [42]. As the number of samples increases, the training time also increases [43]. Over-sampling methods include:

1) *Random Over-Sampling (ROS):* ROS is the earliest over-sampling technique developed, which copies random minority class samples to achieve a more balanced class distribution of samples [44].

2) *Adaptive Synthetic (ADASYN):* This method uses a weighted distribution of the minority class samples based on their difficulty learning. More synthetic samples are generated for minority samples harder to learn than the easier ones [45].

3) *Synthetic Minority Over-Sampling Technique (SMOTE):* Synthetic samples are generated by interpolating k Nearest Neighbors (kNN) of each of the minority samples [46].

4) *Synthetic Minority Over-Sampling Technique - Nominal Continuous (SMOTE-NC):* This is a generalised version of SMOTE that accommodates both continuous and nominal data [46].

5) *Borderline SMOTE:* This method performs SMOTE on borderline samples, which are instances that are often misclassified by their nearest neighbours [47].

6) *Support Vector Machine (SVM) SMOTE:* This method oversamples minority samples along the borderline and uses an SVM classifier for predicting new instances [48].

7) *KMeans SMOTE:* This method uses the combination of KMeans clustering and SMOTE method to form K clusters through clustering and then uses over-sampling to retain clusters that contain many minority samples. These clusters will be allocated to synthetic samples and then put into clusters with insufficient samples in the minority class. Finally, SMOTE balances the proportion of categories in each cluster [49], [50].

The hybrid method is a combination of under-sampling and over-sampling. Under-sampling and over-sampling have unavoidable disadvantages: under-sampling may discard useful information, while over-sampling may lead to overfitting. To break through these limitations, a technique combining under-sampling and over-sampling has been proposed. These include (i) SMOTE-ENN [44], which combines SMOTE for over-sampling and ENN for under-sampling, and (ii) SMOTE-Tomek [44], which uses SMOTE for over-sampling and Tomek links for under-sampling. The purpose of using these two methods is to balance the training dataset and remove the noisy points at the wrong side of the decision boundary, to find better clusters and create models with good generalisation ability.

2) ALGORITHM-LEVEL METHODS

Algorithm-level methods are techniques wherein (i) standard machine learning classifiers are modified and associated with a weight or cost variable, or (ii) the classifier itself is unaffected by the skew distribution [51]. Many scholars have published relevant research results discussing the class-imbalanced problem at the algorithm level [52]–[54].

3) HYBRID SYSTEMS

Hybrid systems involve a combination of sampling techniques and algorithmic methods [55]. They use data-level methods to process data externally and adjust the distribution of categories in the sample. Then algorithms are used internally to modify the learning process [56]. In this way, the model will not skew the majority class too much during classification [9]. The common ensemble methods are as follows:

1) *Balanced Bagging:* This method implements bagging and uses RUS to make the dataset balanced. It resamples each subset of the data before using each integrated estimator. Therefore, its advantage over sci-kit- learn is that it uses two additional parameters that control the behaviour of the random sampler: sampling strategy and replace [57].
2) *Balanced Random Forest:* This method first draws bootstrap samples from the minority class, then randomly draws with replacement the same number of instances from the majority class, creating a balanced sample from which each tree is drawn. The majority vote determines the prediction 58].
3) *Easy Ensemble:* In this method, classifiers are trained on balanced subsets using AdaBoost, then the output of each classifier is combined, creating an ensemble classifier [59].
4) *Random Under-Sampling Boost (RUSBoost):* This method makes sampling and boosting combined and performs RUS in each round of boosting [60].
5) *Balance Cascade:* This method is a double integration algorithm combining bagging and boosting. The iterative method is used to extract a partial subset from the majority class and combine it with the minority class to form a base learner, eliminating the majority class samples that can be correctly classified during training.

This method pays more attention to samples that are easily misclassified [59].

Although both easy ensemble and balanced cascade are called exploratory under-sampling, each time, they extract a subset from the majority class to learn the classifier. But both mainly use AdaBoost to train each bag, and it is classified as ensemble methods [9], [28].

Various ensemble-based resampling techniques, i.e., Balanced Bagging [57], Balanced Random Forest [58], [61], Easy Ensemble [59], RUSBoost [60], and Balance Cascade [59], are widely known. Random balance, SMOTEBoost, and RUS- Boost are identical due to random balance. The randomness and repetition of ensemble methods rely on random balance because each classifier utilizes the random ratio during sample training with different class proportions. SMOTE and RUS balance the samples concerning a minority as well as a majority class. The hybrid method of SMOTE and RUS provided better performance than other state-of-the-art combined ensemble methods such as SMOTEBoost and RUSBoost [62], [63]. The combination of UnderBagging and OverBagging termed as Under-OverBagging based on resampling bagging algorithm has proposed by Qian *et al.*, [64] that oversampled the minority class and undersampled the majority class. The resampling ratio is calculated through the ratio of the minimum class size and the maximum class size.

KNN, naïve Bayes, and neural networks are widely employed as base learners both as homogeneous and heterogeneous ensembles. Previous researches show that the performance of heterogeneous ensembles is highly efficient. Another method developed by Liu and Zhou was named as easy ensemble [59] for data resampling using ensemble methods. Easy ensemble keeps the undersampling method's efficiency higher and reduces the risk of ignoring potentially useful information in majority class samples. It has been observed that using an ensemble as a base classifier is more effective for imbalance classification than using a single classifier. Balance Cascade tries to use guided rather than random deletion of majority class samples. In contrast to Easy Ensemble, it works in a supervised manner. Since Balance Cascade removes correctly classified majority class examples in each iteration, it should be more efficient on highly imbalanced data sets.

Marcelino *et al.* [65] demonstrated that ensemble learners might be affected by the dataset size, an important result since collecting additional data may be costly or infeasible in some cases. Thus, since dataset size may affect classification performance, it is important to examine novel approaches to this problem. Johnson and Khoshgoftaar [17] examined the effects of datasets size and balance levels on the classification performance of various ensemble methods. They concluded that the average AUC value increases within each level of class imbalance as the dataset size increases. Similarly, within each dataset size, the average AUC value increases as the minority distribution increases. In general, ensemble learning methods perform better than any single base learner, tend to be less susceptible to overfitting, and can reduce the bias during data resampling.

RUS [29] is a computationally cheap baseline method that naturally extends to the multi-class case and brings no distortion to class distribution. It is risky because it deletes random samples without checking their potential significance or relevance. TL [38] is a method of border and noise-cleaning. The algorithm is easily extendable to the multi-class case. Still, its computational complexity is higher because it is needed to find the nearest neighbours of each point in the data set. Also, the number of found links is limited because the nearest neighbours will break many candidate pairs from the same class. CNN [39] utilized the one nearest neighbour algorithm to choose which majority sample can be removed. The issue with this method is that it is sensitive to noise by preserving noisy samples. OSS [36] adds the use of TomekLinks to CNN to remove links that are considered noisy. NCR [35] combines C-NN and OSS to remove more noise samples. NM [34] is a binary undersampling algorithm that uses average distances between a given point and the nearest or farthest points of an opposite class. It undersamples only the largest major class because of intrinsic constraints of the binary NearMiss algorithm. NearMiss technique highly distorts a distribution of the major class, also NearMiss-4 has no meaning in the multi-class case.

ROS [44] is a baseline method in which we oversample all minor classes with a random selection of points up to the number of points in the largest major class. However, it can get many instances with the same points, which may not be good for some learning algorithms. ADASYN [45] oversampling algorithm for the multi-class case creates points adaptively to minor classes distributions. The algorithm is not computationally efficient because it computes the nearest neighbours twice, firstly for the whole data to find many points to generate. SMOTE [46], [66] is a widely used multi-class case algorithm. SMOTE has some drawbacks: firstly, its computational complexity is quadratic in the size of the minor class because of the k-nearest neighbour's search. Secondly, selecting target points from the nearest neighbours creates a serious distortion of the minor class distribution. Some points will never be selected as targets; new points are generated as edges of a graph but not in the middle of the distribution. Borderline-SMOTE algorithm [47], [67] creates new points as linear combinations of the borderline minor class points. We have found some drawbacks of the algorithm: 1) low computational efficiency because of k-nearest neighbours to the minor class from the whole data set, 2) a substantial distortion of the minor classes distributions, even more than with pure SMOTE. SMOTE-SVM [48], [68] instead focuses on creating samples on the decision borders of minority and majority classes created by the SVM classifier.

B. PERFORMANCE EVALUATION METRIC FOR IMBALANCED DATA

The performance of a classifier is commonly determined through a confusion matrix shown in Table 2, where True

TABLE 2. Confusion matrix.

Data Class		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Positive (TP) is the number of correctly classified positive instances, False Negative (FN) is the number of positive instances incorrectly classified to be negative. False Positive (FP) is the number of negative instances incorrectly predicted as positive. In contrast, True Negative (TN) is the number of correctly predicted negative instances [69]–[71]. From the confusion matrix, many standard evaluations metrics can be derived [72], [73]. The most commonly used metric is accuracy, given by Eq. 1.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

However, most studies on imbalanced class data point out that accuracy may not be an appropriate metric in imbalanced datasets [74]. This is because, in most applications, the minority class is often more important, requiring methods with improved recognition rates [75], [76], and errors (FN and FP) have varying degrees of consequence. For instance, in cancer diagnosis, one is more interested in correctly detecting the minority (i.e., positive) cases to diagnose and treat the patient effectively. Incorrectly diagnosing a person as cancer-positive could entail additional, unnecessary costs for further medical tests. On the other hand, incorrectly classifying a person as cancer-negative could delay necessary treatment and cost the person's life.

We describe an alternative performance evaluation metric, the area under the Receiver Operating Characteristic curve. The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate ($TPR = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$) on the y-axis against the False Positive Rate ($FPR = \mathrm{FP}/(\mathrm{TN} + \mathrm{FP})$) on the x-axis at various threshold values [77]–[79]. The area under the ROC curve (AUC) identifies the classifier's ability to distinguish between classes and compares ROC curves [80], [81].

C. APPLICATION OF CLASS IMBALANCE METHODS TO CANCER DATASETS

Concerning cancer, a comprehensive review of data-level methods for diagnosing various types of cancer was performed in the research of Sara *et al.* [13]. Compared with other types of cancer, there is less study on class imbalance methods for lung cancer. Few researches also classified the Lung nodules [82], [83], chest-related diseases [84], [85], identification of thoracic diseases [86], forecasting of COVID-19 [83], [87], [88].

III. DATA DESCRIPTION

In this study, we utilise two different lung cancer datasets: (i) the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial and (ii) the National Lung Screening Trial (NLST). As shown in Figure 1, these two datasets are imbalanced in class, and they will be explained below.

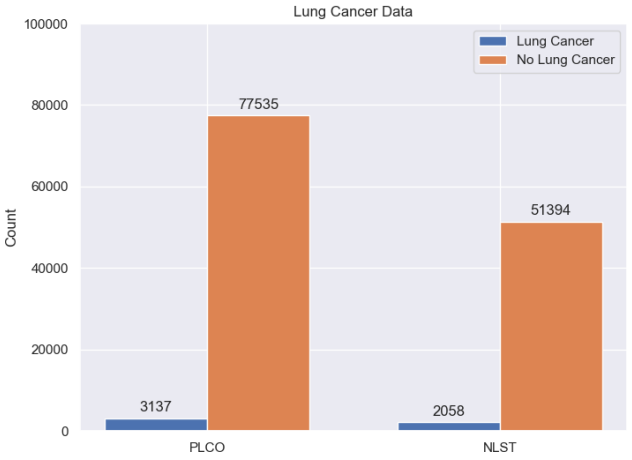


FIGURE 1. PLCO and NLST lung cancer data.

A. PLCO DATASET

The PLCO dataset collects anonymised information of men and women age 55 to 74 years, including their responses to baseline and supplementary questionnaires, smoking status, screening test results, diagnostic and treatment procedures [89]. The initial data consists of 154,897 participants, and after performing data cleaning discussed in Section IV-A.1 and Section IV-A.3, the number of participants was reduced to 80,672. Among them, 3,137 or about 3.89%, have confirmed lung cancer, while the rest have no confirmed lung cancer. We took a subset containing age, Body Mass Index (BMI) value and category, x-ray history, education, smoking status, number of years smoking, pack-years, number of years since quitting smoking, family history of lung cancer, history of bronchitis and emphysema and confirmed lung cancer. These variables were identified in the PLCO model developed to predict lung cancer risk [90].

B. NLST DATASET

The NLST dataset collects participant information to compare Low Dose Computed Tomography (LDCT) with chest radiography in lung cancer screening. The data contained information from 53,452 participants. There are 2,058 participants with confirmed lung cancer or about 3.85% of them.

In this dataset, we created a subset containing variables similar to the first PLCO subset, namely, age, weight, height, x-ray history, education, smoking status, number of years smoking, pack-years, age when participant quit smoking, history of lung cancer of brother, child, father, mother and sister, history of bronchitis and emphysema and confirmed lung cancer.

IV. METHOD

This research is to explore the method of a class-imbalanced dataset in biomedical data. The confirmed lung cancer cases in the PLCO and NLST datasets make up 3.89% and 3.85%

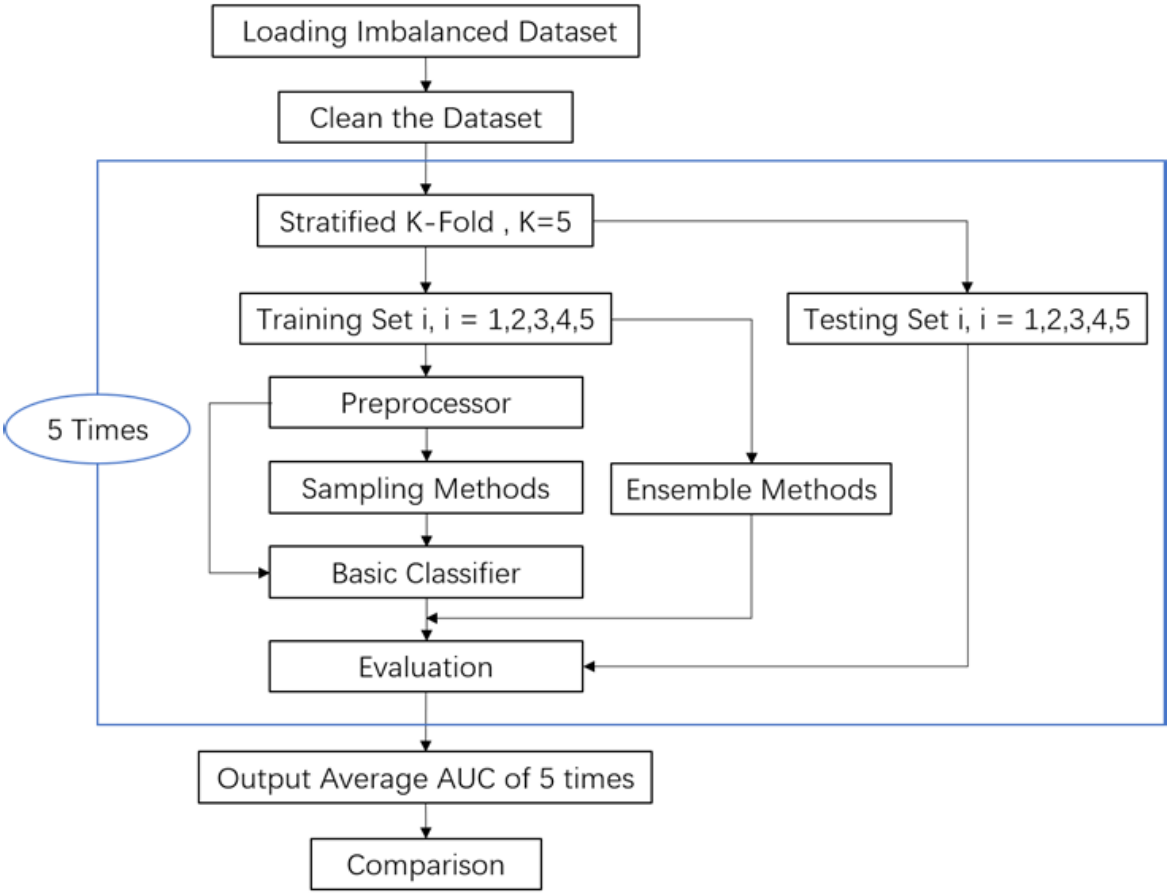


FIGURE 2. Procedure flow.

of the respective populations. This low proportion of positive cases indicate that the class distribution is imbalanced. Therefore, class imbalance techniques are applicable to predict the presence of lung cancer. This research uses three classifiers as baseline models according to the type of class-imbalanced method to be explored. It performs the following two types: (i) perform sampling techniques and build classification models, or (ii) perform ensemble methods. The specific workflow is shown in Figure 2, and this section will explain the methods used in the research.

A. DATA PRE-PROCESSING

Data pre-processing includes addressing the issue of missing values and adjusting the features of the datasets. The part about scaling numerical data and one-hot encoding of categories features will be discussed later.

1) HANDLING MISSING VALUES FOR THE PLCO DATASET

Initial data from the PLCO Lung dataset consists of 154,897 participants. We excluded 4,953 participants with no indicated cigarette smoking status cig stat. Whenever this information was unknown, other variables such as the number of years smoking, pack-years and years since quitting smoking were also unknown. One would not reasonably clean these data without information on whether one is a current, former or never smoker. Variables containing a mixture of categorical and numerical data were cleaned. For instance, the number of years since quitting smoking variable cig stop contained the number of years for some former smokers, zero for current smokers, but had no response for some former smokers and non-smokers. The latter is reflected as NaN and had to be cleaned up. For non-smokers, we set this to be equal to the individual's age (i.e., we assume non-smokers to have ceased smoking since their birth). For those with unknown X-ray history, we set the value to 3, corresponding to the category value that indicates the participant ''does not know'' the answer. For current smokers with an unknown number of years smoking and pack-years, we set their respective cig years and pack-years with the median values for current smokers. Likewise, for former smokers with an unknown number of years smoking, pack-years and years since quitting, we set their respective cig years, pack-years and cig stop with the median value for former smokers. For those with an unknown family history of lung cancer, we set the value to 8, indicating a new category value. For the rest of the variables where we could not reasonably assume values for the cleanup, we used SimpleImputer from scikit-learn [91], [92]. To handle missing values of the numerical

BMI variable, we used the median strategy. In contrast, for categorical variables represented by numbers, namely, education, history of bronchitis and emphysema, we used the most frequent strategy.

We also made a function to map the BMI value BMI curr. We have just imputed to their corresponding categories BMI cure as per World Health Organization (WHO) standard categorisation of BMI. Further, we created a new subset of the cleaned dataset containing our desired features (age, BMI category, x-ray history, education, smoking status, number of years smoking, pack-years, number of years since quitting smoking, family history of lung cancer, history of bronchitis and emphysema), and the target variable (confirmed lung cancer).

2) HANDLING MISSING VALUES FOR THE NLST DATASET

We converted the columns' data type to numeric since they were all initially cast as a string. For the missing height and weight values, we used imputation with the median strategy. We computed the BMI value from the height and weight values and mapped the result to the BMI category using the same mapper we used in PLCO. Current smokers have missing entries for their age when they quit smoking, so we set them to their age. We imputed their median values for former smokers with missing entries for their age when they quit smoking. We then computed the corresponding cig stop value by taking the difference of the participant's age, and age quit to align it with the definition in PLCO data. Lung cancer history of family members in NLST are indicated in separate fields for brother, child, father, mother and sister. For the missing entries in these fields, we used imputation with the most frequent strategy. We then collapsed these features in a single column, lung FH, by taking their resulting logical OR. For the missing history of bronchitis and emphysema, we used imputation with the most frequent strategy. We also introduced the binary target variable confirmed lung with a value of 1 if the participant has confirmed lung cancer and 0 otherwise, based on the variable conflict. It simplifies our study to a binary classification problem.

Further, we created a new subset of the cleaned NLST dataset containing our desired features (age, BMI category, x-ray history, education, smoking status, number of years smoking, pack-years, number of years since quitting smoking, family history of lung cancer, history of bronchitis and emphysema), and the target variable (confirmed lung cancer), using the same order and exact column names as the PLCO dataset.

3) MAKE PLCO AND NLST DATASET CONSISTENT

In this section, the two datasets after preliminary cleaning are further processed, and it is expected that the characteristics of the two datasets are consistent. We removed the PLCO non-smokers from the dataset because the NLST excludes non-smokers from their screening selection criteria. We also changed the PLCO's former smokers cig stat with a value of 2 to 0 to align with NLST's former smoker's cigsmok value of 0. NLST's categories 8, 95, 98 and 99 did not correspond to PLCO's education categories for the education feature. We calculated the mode for NLST's education variable EDUCAT, which was 3, and used this value instead for the mentioned categories. Family history of lung cancer in PLCO had categories 8 and 9, which did not correspond to NLST's corresponding categories. We used the PLCO's mode for lung fh, which was 0, for these categories. For x-ray history, to align with NLST's binary 0-1 values, we collapsed PLCO's ''Yes, Once'' and ''Yes, More Than Once'' (with values 1 and 2, respectively) into the same value of 1. Also, for those who answered ''Do not Know'' (with the value of 3), we assumed that if they were not sure of their x-ray history, the results would not have been available, so we set those at 0. Finally, we renamed NLST's feature names to follow those of PLCO's for easier reference. We identified the following variables as categorical: BMI curr, bronchit f, cig stat, EDUCAT, emphys f, lung FH, Xray history, while the following variables are numerical: age, cig stop, cig years and pack years.

B. SPLIT DATASET

The researcher used Stratified KFold (K = 5) to split the dataset, dividing the entire development set into five disjoint subsets while still maintaining the sample category ratio. This method uses four-fifths of the dataset for each split. As the training set, the remaining one-fifth is used as the test set. Each split can be regarded as the *i*th time (i = 1, . . . , 5), and AUC is calculated on the *i*th test set [93]. It is worth noting that the test set obtained each time will be placed aside, and it will not participate in any stage of scaling or recoding and model building. Since the over-sampling method will copy or synthesise some minority samples, the data obtained in this way cannot represent the original dataset, so the test set should be far from the training process.

C. SCALING AND ENCODING DATA

Scaling data and re-encoding should be applied before sampling because some sampling methods are related to the distance between the data. For example, All-KNN is based on the Euclidean distance of the data, and the magnitude of the excessive difference will affect the sampling effect. The methods of scaling and encoding will be explained in detail.

1) FEATURE SCALING FOR NUMERIC DATA

As part of data pre-processing, we transformed the numeric data to a range of [0,1] using Eq. 2.

$$X' = (x - \min(x))/(\max(x) - \min(x)) \tag{2}$$

2) ONE-HOT ENCODING FOR CATEGORICAL DATA

We performed one-hot encoding for categorical data. Each categorical feature with n categories is converted to n binary (0-1) features [94], [95].

D. CLASS-IMBALANCED METHODS

The class-imbalanced learning methods used in this research mainly include data-level methods and hybrid systems (this research mainly explores the imbalance technologies in the Imblearn library). We mainly use resampling techniques for data-level methods, including under-sampling, over-sampling, and hybrid sampling methods. Under-sampling methods: Random Under Sampling (RUS), All k-Nearest Neighbors (All k-NN), Cluster Centroids (CC), Edited Nearest Neighbors (ENN), Instance Hardness Threshold (IHT), Near Miss (NM), Neighbourhood Cleaning Rule (NCR), One-Sided Selection (OSS), Repeated ENN (RENN), Tomek Links (TL). In this study, due to the huge dataset, Condense Nearest Neighbors(CNN) is an al- algorithm based on 1-NN, which requires much time to run. Therefore, CNN is not discussed in this article. Over-sampling methods: Random Over Sampling (ROS), Adaptive Synthetic (ADASYN), Synthetic Minority Over-sampling Technique (SMOTE), SMOTE-Nominal Continuous (SMOTE-NC), Borderline SMOTE, Support Vector Machine (SVM) SMOTE, KMeans SMOTE. Hybrid sampling methods: SMOTE-ENN, SMOTE-Tomek. Those data-level methods are combined with classifiers to predict lung cancer cases. For Hybrid systems, we trained them with the inherent classifier. They are Balanced Bagging, Balanced Random Forest, Easy Ensemble, and Random Under-Sampling Boost (RUSBoost). The Balance Cascade algorithm has been continuously adjusted by the Imblearn library in recent years and was finally abandoned in version 0.6.0, so this article will not discuss this method.

E. BUILDING CLASSIFIERS

This study uses three classic classifiers as the baseline model to find the most suitable class-imbalanced technique for the dataset based on this standard: (i) Logistic Regression (LR), (ii) Random Forest (RF), and (iii) Linear Support Vector Classification (Linear SVC).

F. EVALUATION

1) EVALUATE SAMPLING – IMBALANCE RATIO

The imbalance ratio (IR) is an essential parameter in imbalanced learning. It measures the proportional relationship between the majority and minority classes in the experiment [96]. The formula is given by Eq. 3:

$$IR = Instances_{Majority}/Instances_{Minority} \tag{3}$$

Most of the data-level methods used in the research are by resampling the majority class or minority class in the original dataset, thereby increasing the minority class samples or reducing majority class samples. Sampling will cause the imbalance ratio of the dataset to change. As IR becomes larger, the disparity in sample size between the majority class and the minority class becomes more significant [97], [98]. The dataset at this time is imbalanced. When the IR value is closer to 1, the dataset tends to be more balanced. Therefore, this paper will use IR to evaluate sampling techniques.

2) EVALUATE MODEL – AUC

This study selected widely-used AUC as the metric to evaluate the ability of each classifier to distinguish between confirmed and no confirmed lung cancer cases. After *i*th attempts, we can get the mean AUC of *i*th training on the *i*th test set. To make the experimental results more accurate and reliable, this study repeated the above process five times and calculated the final mean AUC to measure the model's predictive ability. In addition, this study will compare the experimental results in the PLCO and NLST datasets and discuss the methods of dealing with class-imbalanced data.

V. RESULTS

This section will list the imbalance ratio provided by the resampling technique and then show the prediction results of the imbalance technique model, which can help analyse the effect of the imbalance technique comprehensively. We have used the area under the curve (AUC) for the evaluation of proposed methods. The AUC performs best when the dataset is imbalanced [10], [69]. Our study had 16 imbalance datasets, so various studies [57], [99], [100] employed the AUC curve as a performance evaluation measure.

A. RESULTS FOR PLCO DATASET

The class-imbalanced PLCO dataset has an imbalanced ratio of 24.7. Through resampling technology, the class proportion of the dataset has changed. Table 3 lists the class distribution in the training set after each sampling.

TABLE 3. Class distribution for data-level methods _-PLCO.

Method	Imbalance Ratio	Majority Samples	Minority Samples
Baseline	24.72	62028.00	2509.60
RUS	1.00	2509.60	2509.60
AllKNN	21.70	54451.40	2509.60
CC	1.00	2509.60	2509.60
ENN	22.31	55987.00	2509.60
IHT	10.54	26440.20	2509.60
NM	1.00	2509.60	2509.60
NCR	22.28	55902.80	2509.60
OSS	24.35	61108.80	2509.60
RENN	20.11	50472.40	2509.60
TL	24.39	61197.80	2509.60
ROS	1.00	62028.00	62028.00
ADASYN	1.00	62028.00	61878.00
SMOTE	1.00	62028.00	62028.00
SMOTENC	1.00	62028.00	62028.00
BSMOTE	1.00	62028.00	62028.00
SVMSMOTE	1.74	62028.00	35636.60
KmeansSMOTE	1.00	62028.00	62031.24
SMOTEENN	0.80	46909.44	58881.80
SMOTETomek	1.00	61849.88	61849.88

Since the sampling occurs in the training set, the baseline of the dataset is the number of samples in the training set (four-fifths of the whole dataset, which is 64537.6). It can be seen from the result that under-sampling changes the majority of samples, over-sampling only processes the minority samples, and the hybrid method changes both categories.

All sampling methods reduce the IR value, and the IR values of over-sampling and hybrid sampling are close to 1, which means that they achieve the class-balanced of the dataset as much as possible.

Applying various under-sampling methods for the PLCO dataset, we show the resulting AUCs for three different classifiers in Table 4. Each classifier had another best under-sampling method. Logistic regression using RUS and Linear SVC had higher scores, 0.7124 and 0.7126, respectively. However, the random forest model using Repeated ENN got the highest mean AUC of 0.8968 in the model using the under-sampling method.

TABLE 4. AUC results for under-sampling methods - PLCO.

Method	Logistic Regression (AUC)	Random Forest(AUC)	Linear SVC(AUC)
Baseline	0.5001	0.8532	0.5000
RUS	**0.7124**	0.8120	**0.7126**
AllKNN	0.5041	0.8926	0.5000
CC	0.6616	0.6809	0.6615
ENN	0.5022	0.8804	0.5000
IHT	0.6755	0.8590	0.6643
NM	0.4745	0.5035	0.4699
NCR	0.5020	0.8783	0.5000
OSS	0.5001	0.8543	0.5000
RENN	0.5016	**0.8968**	0.5000
TL	0.5001	0.8542	0.5000

For over-sampling methods, ROS had the best performance among the three classifiers. These are shown in Table 5. The random forest had the highest mean AUC of 0.8994 among them.

TABLE 5. AUC results for over-sampling methods - PLCO.

Method	Logistic Regression (AUC)	Random Forest (AUC)	Linear SVC (AUC)
Baseline	0.5001	0.8532	0.5000
ROS	**0.7129**	**0.8994**	**0.7130**
ADASYN	0.7124	0.8706	0.7124
SMOTE	0.7109	0.8693	0.7113
SMOTENC	0.7027	0.8835	0.7032
BSMOTE	0.7086	0.8703	0.7085
SVMSMOTE	0.6773	0.8677	0.6747
KmeansSMOTE	0.6758	0.8642	0.6675

For Hybrid Methods shown in Table 6, SMOTEENN achieved a higher mean AUC in logistic regression and Linear SVC. Nevertheless, using SMOTETomek with logistic regression had a higher mean AUC of 0.8684.

TABLE 6. AUC results for hybrid methods - PLCO.

Method	Logistic Regression (AUC)	Random Forest (AUC)	Linear SVC (AUC)
Baseline	0.5001	0.8532	0.5000
SMOTEENN	**0.7134**	0.8583	**0.7126**
SMOTETomek	0.7107	**0.8684**	0.7116

For ensemble methods shown in Table 7, balanced bagging achieved the highest mean AUC, followed by balanced random forest.

TABLE 7. AUC results for ensemble methods - PLCO.

Method	AUC
Balanced Bagging (BB)	**0.8403**
Balanced Random Forest (BRF)	0.8143
EasyEnsemble	0.7188
RUSBoost	0.7159

The researchers measured all resampling methods in the random forest model with the highest baseline value. In Figure 3, yellow represents the baseline, green represents the under-sampling methods, orange represents the over-sampling methods, and blue represents the hybrid methods. The baseline AUC value in PLCO is 0.8532; it can be seen that the lowest value that appears in Near Miss is 0.5035, the highest value appears in ROS, and its AUC value is 0.8994. Observing the bar chart shows that the AUC displayed by the under-sampling method has more significant fluctuations than other methods. Through calculation, the standard deviation (SD) of under-sampling in PLCO is 0.1251, and the SD value of over-sampling is 0.0123. There are only two-hybrid methods, so their SD is not calculated. Also, we separately calculated the standard deviation of ensemble methods (because this method is a separate classifier) as 0.0643. The result is between over-sampling and under-sampling. It shows that over-sampling is more stable than other imbalanced learning, and under-sampling is the most unstable. Among all the class imbalance techniques tested in the PLCO dataset, random forest using ROS performs best.

B. RESULTS FOR NLST DATASET

The NLST dataset is also an extremely imbalanced dataset, with an imbalance rate of 25.2. The imbalance rate of the dataset obtained by the sampling method is shown in Table 8. We can see similar results to the PLCO dataset. Over-sampling and hybrid sampling make the IR adjustment of the

TABLE 8. Class distribution for data-level methods - NLST.

Method	Imbalance Ratio	Majority Samples	Minority Samples
Baseline	25.22	62056.00	2460.80
RUS	1.00	2460.80	2460.80
AllKNN	23.48	57784.80	2460.80
CC	1.00	2460.80	2460.80
ENN	23.76	58457.00	2460.80
IHT	13.98	34400.12	2460.80
NM	1.00	2460.80	2460.80
NCR	23.69	58288.00	2460.80
OSS	25.10	61776.24	2460.80
RENN	22.37	55055.80	2460.80
TL	25.14	61865.40	2460.80
ADASYN	1.00	62056.00	62037.80
ROS	1.00	62056.00	62056.00
SMOTE	1.00	62056.00	62056.00
SMOTENC	1.00	62056.00	62056.00
BSMOTE	1.00	62056.00	62056.00
SVMSMOTE	1.00	62056.00	62056.00
KmeansSMOTE	1.00	62056.00	62059.12
SMOTEENN	0.91	54162.72	59207.16
SMOTETomek	1.00	62007.36	62007.36

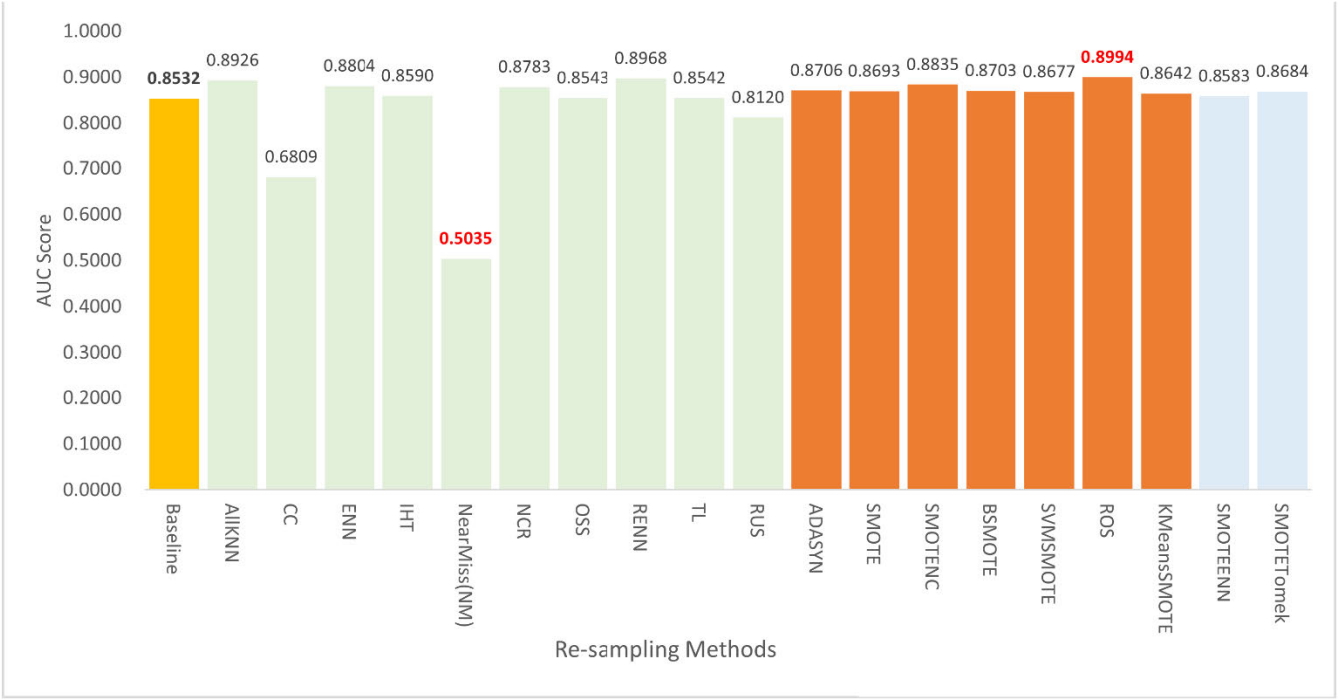


FIGURE 3. Comparison of sampling method on random forest in PLCO.

dataset approximately 1. The sample size in the training set shows that the number of samples is reduced after using the under-sampling technique. In contrast, the total number of samples is higher than the original dataset after using other methods.

Table 9 shows the resulting AUCs upon applying various under-sampling methods in conjunction with three different classifiers for the NLST dataset. Each classifier had another best under-sampling method. However, for Logistic regression and linear SVC, the difference between the best performing AUC is very small, and their sampling methods are both RUS. Besides, the performance of Random Forest using Repeated ENN is much better than other models in under-sampling methods.

TABLE 9. AUC results for under-sampling methods - NLST.

Method	Logistic Regression(AUC)	Random Forest(AUC)	Linear SVC(AUC)
Baseline	0.5000	0.8502	0.5000
RUS	**0.6528**	0.8323	**0.6543**
AllKNN	0.5000	0.8812	0.5000
CC	0.5907	0.6737	0.5889
ENN	0.5000	0.8799	0.5000
IHT	0.5088	0.8860	0.5000
NM	0.5084	0.5477	0.5191
NCR	0.5000	0.8702	0.5000
OSS	0.5000	0.8500	0.5000
RENN	0.5000	**0.8930**	0.5000
TL	0.5000	0.8501	0.5000

We show the AUC results for the over-sampling methods in Table 10. Logistic regression is similar to the best over-sampling method of Linear SVC. Random forest with ROS achieved the highest mean AUC of 0.8960.

TABLE 10. AUC results for over-sampling methods - NLST.

Method	Logistic Regression(AUC)	Random Forest(AUC)	Linear SVC(AUC)
Baseline	0.5000	0.8502	0.5000
ROS	0.6553	**0.8960**	0.6539
ADASYN	0.6526	0.8802	0.6531
SMOTE	0.6543	0.8799	**0.6544**
SMOTENC	0.6544	0.8774	0.6543
BSMOTE	**0.6562**	0.8790	0.5000
SVMSMOTE	0.6378	0.8783	0.6378
KmeansSMOTE	0.5829	0.8654	0.5967

For hybrid methods shown in Table 11, SMOTETomek achieved a higher mean AUC than SMOTEENN for all three classifiers in the NLST dataset.

TABLE 11. AUC results for hybrid methods - NLST.

Method	Logistic Regression (AUC)	Random Forest (AUC)	Linear SVC(AUC)
Baseline	0.5000	0.8502	0.5000
SMOTEENN	0.6549	0.8588	0.6543
SMOTETomek	**0.6550**	**0.8800**	**0.6550**

AUCs of ensemble methods performed in the NLST dataset are shown in Table 12. Balanced bagging achieved the highest mean AUC, followed by balanced random forest.

Similarly, like the PLCO dataset, we measure the performance of the sampling method in the random forest, as shown

TABLE 12. AUC results for methods - NLST.

Method	AUC
Balanced Bagging (BB)	**0.8588**
Balanced Random Forest(BRF)	0.8476
Easy Ensemble	0.6606
RUSBoost	0.6567

in the figure. It can be seen that the AUC value of the under-sampling Near Miss is the lowest, and the AUC value of the over-sampling ROS is the highest. By calculating the AUC standard deviation of various sampling methods in the NLST dataset, the SD value of under-sampling is 0.1140, and the SD value of over-sampling is 0.0089. In addition, the standard deviation of hybrid systems is 0.1124, which is between over-sampling and under-sampling. Combining the standard deviation performance and the AUC in each method, under-sampling fluctuates wildly compared to over-sampling, which is more stable.

In general, AUCs obtained in the NLST dataset have been lower than the AUCs obtained in the PLCO dataset, indicating an inherent difference in the data.

VI. DISCUSSION

In this section, we will discuss the application of class-imbalanced technology in this study in two aspects. One is to discuss different class-imbalanced techniques, and the other is to combine the performance of the two datasets to analyse the results.

A. THE EFFECTS OF IMBALANCED LEARNING

Each classifier is combined with different imbalance techniques in this study, including data-level over-sampling, under-sampling, hybrid method, and methods. Among the three baseline classifiers, the mean value of the random forest is much higher than logistic regression and Linear SVC, and random forest models provide the highest mean value of AUC with different sampling techniques. It shows that the random forest classifier is suitable for these imbalanced medical data used in this study. It is worth noting that although the baseline AUC values of logistic regression and Linear SVC are as low as 0.5, the AUC values of most models have been significantly improved through the use of class imbalance techniques. It shows that the class imbalance technique helps to enhance the ability of model classification. Besides, most of the average AUC in over-sampling methods is higher than other sampling methods. The results show that the over-sampling way is suitable for the imbalanced medical data used in this study. The following will discuss the class imbalance learning in two aspects: the class ratio (IR value) of the samples generated from resampling and the stability of the class imbalance techniques.

It is worth noting that although the baseline AUC values of logistic regression and Linear SVC are as low as 0.5, the AUC values of most models have been significantly improved through the use of class imbalance techniques. It shows that the class imbalance technique helps to enhance the ability of model classification. Besides, most of the average AUC in over-sampling methods is higher than other sampling methods. The results show that the over-sampling way is suitable for the imbalanced medical data used in this study. The following will discuss the class imbalance learning in two aspects: the class ratio (IR value) of the samples generated from resampling and the stability of the class imbalance techniques.

To explore the relationship between the imbalance method and the model's AUC, we use IR to measure the ability of resampling technology to adjust the class distribution. From the sampling results, under-sampling discards part of the majority samples, over-sampling duplicates or synthesises minority samples, and the composite method samples all classes. However, in this study's extremely imbalanced dataset, the performance of under-sampling is not excellent, and the IR of most under-sampling methods is very high. Because under-sampling needs to discard many majority class samples to balance with the minority class, this is likely to lose valuable information. When observing the over-sampling and hybrid methods that perform well after combining with the classifier, the researchers found that the minority class samples were significantly increased. The samples were more than the original dataset, and their IR values were all-around 1. Therefore, it can be considered that the resampling method can adjust the sample distribution of the sample to make the IR of the dataset close to 1, which is beneficial to improve the model's predictive ability. Besides, the researchers also used the standard deviation to assess the stability of the imbalanced learning technique. Since the performance of the random forest classifier is better than other baseline classifiers, the researchers exemplified the AUC value of the resampling model used in the random forest. By calculating the standard deviation (SD value) within each type of resampling method, the SD value of methods (hybrid systems) is also calculated separately. We get the highest SD value of under-sampling (In PLCO: 0.1251; In NLST: 0.1140) and the lower SD value of over-sampling (In PLCO: 0.0123; In NLST: 0.0089).

It shows that different methods may have very different results when under-sampling is used, and using different over-sampling methods may get relatively similar results. The standard deviation of over-sampling is much smaller than under-sampling, indicating that the over-sampling method is stable. Therefore, if the resampling method is used to process extremely imbalanced datasets like this research, over-sampling is recommended. Because the over-sampling method is relatively stable, it will not produce significant results due to selecting different methods.

B. EVALUATION OF IMBALANCED LEARNING TECHNIQUES APPLIED TO THE TWO DATASETS

After comparing the performance of different imbalance methods in the two datasets, similar results can be obtained: under-sampling pre-processing the two datasets, RUS has

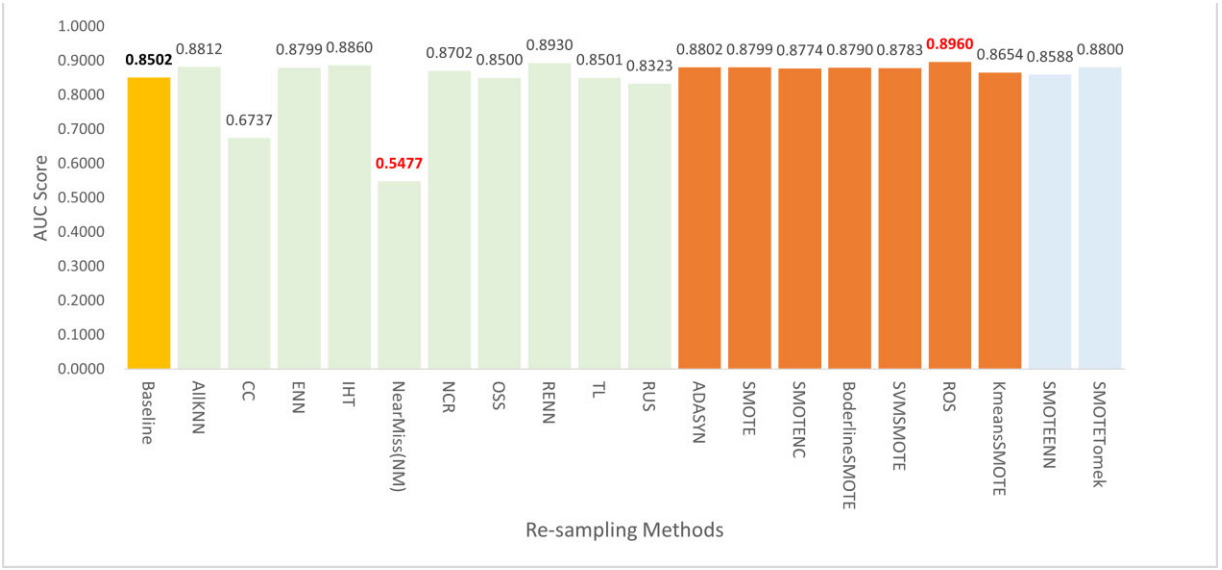


FIGURE 4. Comparison of sampling method on random forest in NLST.

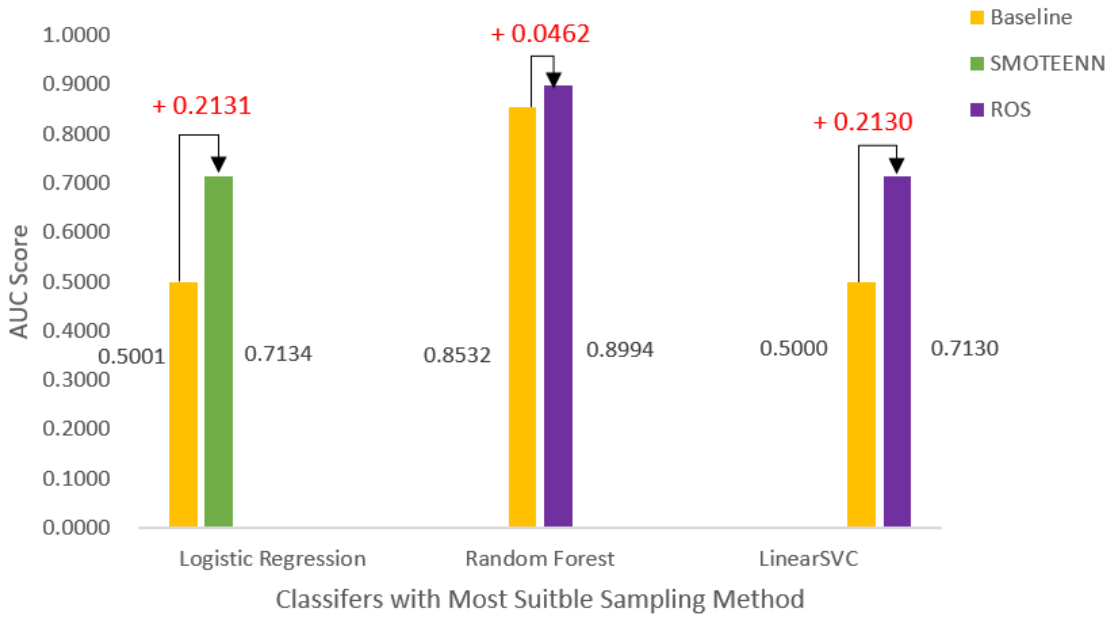


FIGURE 5. Comparison of best performing sampling methods against classifier baseline on PLCO.

shown good logistic regression and linear SVC performance. The combination of Repeated ENN and random forest both got the highest average AUC in under-sampling. In the example of using the over-sampling technique, the random forest combined with ROS performed best among all models in both datasets. For ensemble methods, a balanced bagging classifier performed well for both datasets.

In Figure 5 and Figure 6, we summarise the best performing sampling methods for each classifier on the two datasets and compare them with the baseline AUC (i.e., no sampling performed). After each classifier is processed by the sampling method in the table, the AUC of the model has been significantly increased. Except for Linear SVC, the best sampling methods for the other two classifiers are ROS, and the performance of ROS in Linear SVC is similar to the best results. Therefore, the random forest model using ROS is more suitable for processing such imbalanced medical datasets and achieving the highest AUC. The Near Miss of the under-sampling method obtained results lower than the AUC value of the corresponding baseline classifier

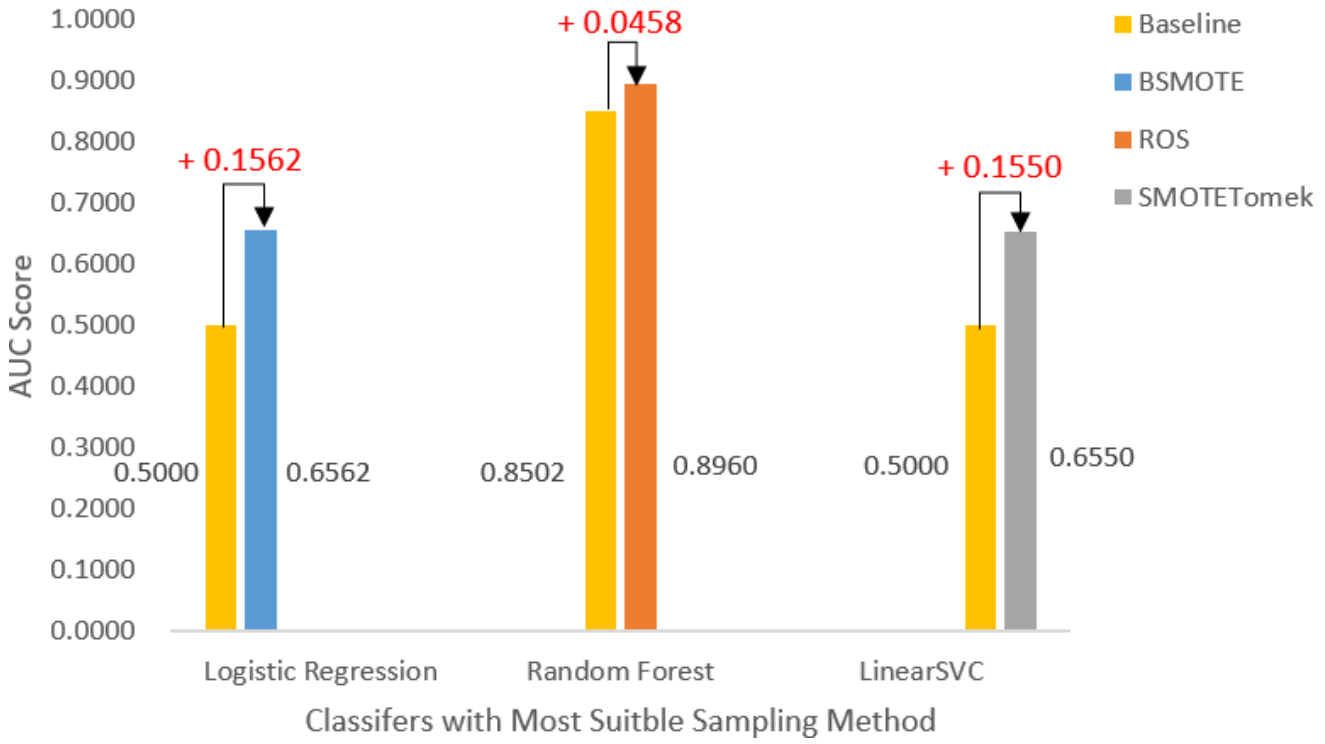


FIGURE 6. Comparison of best performing sampling methods against classifier baseline on NLST.

in both datasets. It performed the worst among all resampling techniques. Therefore, the AUC values obtained by Near Miss on the three classifiers are all the lowest, and it can be considered that it is not suitable for the datasets with an imbalance rate of about 25 used in this study.

Conversely, the random forest model that uses ROS as a whole is more suitable for the highly imbalanced lung cancer dataset used in the research and can achieve the highest AUC. The difference is that SMOTETomek performs very well in the NLST dataset in hybrid methods. The average performance of SMOTEENN in the PLCO dataset is slightly higher than that in the NLST dataset. It shows that there are still some potential differences between the two datasets.

It may be worthwhile to include algorithm-level methods to complete the suite of class imbalance techniques and evaluate their predictive performance. However, the costs and weights assigned to the algorithm-level methods must be as close as possible to realistic values.

VII. CONCLUSION

In this study, we have investigated class imbalance techniques, including data-level and hybrid systems, to predict the presence of lung cancer. Two medical datasets related to lung cancer (PLCO and NLST) with imbalance ratios of 24.7 and 25.2 are used in this research. The imbalanced learning method is used to solve the problem of a skewed majority in prediction. This research discusses 23 imbalanced learning methods, including ten under-sampling techniques, seven over-sampling techniques, two-hybrid resampling methods, and four hybrid systems. The class imbalance technology adjusts the majority or minority samples by discarding the majority samples, copying or synthesising the minority samples to balance the categories in the dataset. In addition, three classic classifiers (logistic regression, random forest, linear SVC) combined with resampling techniques were used to train the dataset. The prediction results obtained using the classifier training pre-processing data (except for null values, etc.) are used as a baseline for comparison with models built using imbalance techniques. The method used to evaluate the sampling technique is the imbalance ratio, and the index used to assess the classification ability of the model is AUC.

Further, the standard deviation was used to measure the stability of class imbalance techniques. This study shows that using the class-imbalance technique has higher performance than the baseline model. Class imbalance technology helps to improve the prediction performance of the model. The data-level technology adjusts the IR of the dataset to be close to 1 through resampling. Among the imbalanced learning methods studied in this paper, the over-sampling technique performed best, and the IR value of the over-sampling dataset was about 1. Most of the models that use over-sampling have higher AUC values than other models. The over-sampling method has higher stability than other methods, and the under-sampling method has the worst stability. Also, the random forest with random over-sampling is the best predictive model, and it is more suitable for the PLCO and NLST datasets related to lung cancer. Using ROS technology to process these two datasets in the random forest model can achieve the highest AUC value.

Conversely, the random forest using Near Miss is even far below the baseline value. Therefore, the combination of ROS technology and the random forest is worthy of promotion. However, there are still some small gaps within different datasets, and compound systems and over-sampling can be suggested to deal with extremely imbalanced biomedical datasets similar to those in the research. The contribution of this research is to prove that the class imbalance techniques can be used to diagnose lung cancer. The over-sampling technique is better than other imbalanced learning methods. Finally, the researchers proposed a model combining ROS and random forest to screen for lung cancer so that more people can receive timely treatment and reduce the loss caused by misdiagnosis. In future research, the new class imbalance technology is worthy of application and exploration. Combining more diverse classifiers and imbalance techniques to achieve higher model prediction capabilities is also worth looking forward. Furthermore, a deep learning-based model, i.e., GNN, AlexNet, ResNet etc., can also be deployed for the imbalance dataset problem.

REFERENCES

[1] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning From Imbalanced Data Sets*. Berlin, Germany: Springer, 2018.

[2] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," in *Proc. 4th Int. Conf. Natural Comput.*, vol. 4, 2008, pp. 192–201.

[3] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[4] 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.

[5] G. M. Weiss and F. Provost, "The effect of class distribution on classifier learning: An empirical study," Rutgers Univ., New Brunswick, NJ, USA, Tech. Rep. 991031549986704646, 2001.

[6] D.-C. Li, C.-W. Liu, and S. C. Hu, "A learning method for the class imbalance problem with medical data sets," *Comput. Biol. Med.*, vol. 40, no. 5, pp. 509–518, 2010.

[7] M. M. Oken, W. G. Hocking, P. A. Kvale, G. L. Andriole, S. S. Buys, T. R. Church, E. D. Crawford, M. N. Fouad, C. Isaacs, D. J. Reding, and J. L. Weissfeld, "Screening by chest radiograph and lung cancer mortality: The Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized trial," *Jama*, vol. 306, no. 17, pp. 1865–1873, 2011.

[8] M. C. Tammemaegi, T. R. Church, W. G. Hocking, G. A. Silvestri, P. A. Kvale, T. L. Riley, J. Commins, and C. D. Berg, "Evaluation of the lung cancer risks at which to screen ever-and never-smokers: Screening rules applied to the PLCO and NLST cohorts," *PLoS Med.*, vol. 11, no. 12, 2014, Art. no. e1001764.

[9] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst., Man, C, Appl. Rev.*, vol. 42, no. 4, pp. 463–484, Jul. 2011.

[10] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249–259, Oct. 2017.

[11] Q. Li and Y. Mao, "A review of boosting methods for imbalanced data classification," *Pattern Anal. Appl.*, vol. 17, no. 4, pp. 679–693, Nov. 2014.

[12] H. He and Y. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*. NJ, USA, 2013.

[13] S. Fotouhi, S. Asadi, and M. W. Kattan, "A comprehensive data level analysis for cancer diagnosis on imbalanced data," *J. Biomed. Informat.*, vol. 90, Feb. 2019, Art. no. 103089.

[14] H. Yu, J. Ni, Y. Dan, and S. Xu, "Mining and integrating reliable decision rules for imbalanced cancer gene expression data sets," *Tsinghua Sci. Technol.*, vol. 17, no. 2, pp. 666–673, Dec. 2012.

[15] J. Diz, G. Marreiros, and A. Freitas, "Applying data mining techniques to improve breast cancer diagnosis," *J. Med. Syst.*, vol. 40, no. 9, pp. 1–7, Sep. 2016.

[16] S. Santiso, A. Casillas, and A. Pérez, "The class imbalance problem detecting adverse drug reactions in electronic health records," *Health Informat. J.*, vol. 25, no. 4, pp. 1768–1778, Dec. 2019.

[17] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, pp. 1–54, Dec. 2019.

[18] V. García, J. S. Sánchez, and R. A. Mollineda, "Exploring the performance of resampling strategies for the class imbalance problem," in *Proc. Int. Conf. Ind., Eng. Appl. Appl. Intell. Syst.* Berlin, Germany: Springer, 2010, pp. 541–549.

[19] R. Longadge and S. Dongre, "Class imbalance problem in data mining review," 2013, *arXiv:1305.1707*. [Online]. Available: http://arxiv.org/abs/1305.1707

[20] J. Liu, Q. Hu, and D. Yu, "A comparative study on rough set based class imbalance learning," *Knowl.-Based Syst.*, vol. 21, no. 8, pp. 753–763, Dec. 2008.

[21] P. Lee, "Resampling methods improve the predictive power of modeling in class-imbalanced datasets," *Int. J. Environ. Res. Public Health*, vol. 11, no. 9, pp. 9776–9789, Sep. 2014.

[22] T. M. Alam, M. A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T. I. Baig, A. Hussain, M. A. Malik, M. M. Raza, S. Ibrar, and Z. Abbas, "A model for early prediction of diabetes," *Inform. Med. Unlocked*, vol. 16, Jan. 2019, Art. no. 100204.

[23] H. M. Nguyen, E. W. Cooper, and K. Kamei, "A comparative study on sampling techniques for handling class imbalance in streaming data," in *Proc. 6th Int. Conf. Soft Comput. Intell. Syst., 13th Int. Symp. Adv. Intell. Syst.*, Nov. 2012, pp. 1762–1767.

[24] E. Burnaev, P. Erofeev, and A. Papanov, "Influence of resampling on accuracy of imbalanced classification," in *Proc. 8th Int. Conf. Mach. Vis. (ICMV)*, vol. 9875, Dec. 2015, Art. no. 987521.

[25] A. C. Liu, "The effect of oversampling and undersampling on classifying imbalanced text datasets," Ph.D. dissertation, Dept. Graduate School, Univ. Texas Austin, Austin, TX, USA, 2004, p. 67.

[26] S.-J. Yen and Y.-S. Lee, "Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset," in *Intelligent Control and Automation*. Berlin, Germany: Springer, 2006, pp. 731–740.

[27] W. Jindaluang, V. Chouvatut, and S. Kantabutra, "Under-sampling by algorithm with performance guaranteed for class-imbalance problem," in *Proc. Int. Comput. Sci. Eng. Conf. (ICSEC)*, Jul. 2014, pp. 215–221.

[28] T. M. Alam, K. Shaukat, I. A. Hameed, W. A. Khan, M. U. Sarwar, F. Iqbal, and S. Luo, "A novel framework for prognostic factors identification of malignant mesothelioma through association rule mining," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102726, doi: 10.1016/j.bspc.2021.102726.

[29] J. Prusa, T. M. Khoshgoftaar, D. J. Dittman, and A. Napolitano, "Using random undersampling to alleviate class imbalance on tweet sentiment data," in *Proc. IEEE Int. Conf. Inf. Reuse Integr.*, Aug. 2015, pp. 197–202.

[30] I. Tomek, "An experiment with the edited nearest-nieghbor rule," *IEEE Trans. Syst., Man, Cybern.*, 1976.

[31] K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed, and M. Xu, "A survey on machine learning techniques for cyber security in the last decade," *IEEE Access*, vol. 8, pp. 222310–222354, 2020.

[32] A. Z. Broder, A. M. Bruckstein, and J. Koplowitz, "On the performance of edited nearest neighbor rules in high dimensions," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, no. 1, pp. 136–139, Jan. 1985.

[33] M. R. Smith, T. Martinez, and C. Giraud-Carrier, "An instance level analysis of data complexity," *Mach. Learn.*, vol. 95, no. 2, pp. 225–256, 2014.

[34] I. Mani and I. Zhang, "kNN approach to unbalanced data distributions: A case study involving information extraction," in *Proc. Workshop Learn. Imbalanced Datasets*, 2003, vol. 126.

[35] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution," in *Proc. Conf. Artif. Intell. Med. Eur.* Berlin, Germany: Springer, 2001, pp. 63–66.

[36] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proc. ICML*, vol. 97, 1997, pp. 179–186.

[37] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-2, no. 3, pp. 408–421, Jul. 1972.

[38] I. Tomek, "Two modifications of CNN," *IEEE Trans. Syst., Man, Cybern.*, 1976.

[39] P. Hart, "The condensed nearest neighbor rule (Corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 3, pp. 515–516, May 1968.

[40] A. I. Marqués, V. García, and J. S. Sánchez, "On the suitability of resampling techniques for the class imbalance problem in credit scoring," *J. Oper. Res. Soc.*, vol. 64, no. 7, pp. 1060–1070, Jul. 2013.

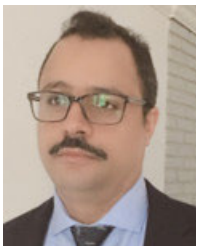
[41] A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Hawalah, and A. Hussain, "Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study," *IEEE Access*, vol. 4, pp. 7940–7957, 2016.
[42] Z. Zheng, Y. Cai, and Y. Li, "Oversampling method for imbalanced classification," *Comput. Inform.*, vol. 34, no. 5, pp. 1017–1037, 2015.
[43] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 2, no. 4, pp. 42–47, 2012.
[44] G. E. Batista, R. C. Prati, and M. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 20–29, 2004.
[45] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Jun. 2008, pp. 1322–1328.
[46] 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, no. 1, pp. 321–357, 2002.
[47] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Proc. Int. Conf. Intell. Comput.* Berlin, Germany: Springer, 2005, pp. 878–887.
[48] H. M. Nguyen, E. W. Cooper, and K. Kamei, "Borderline over-sampling for imbalanced data classification," *Int. J. Knowl. Eng. Soft Data Paradigms*, vol. 3, no. 1, pp. 4–21, Apr. 2011.
[49] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE untuk menangani ketidakseimbangan kelas dalam klasifikasi penyakit diabetes dengan C4. 5, SVM, dan naive Bayes," *J. Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 89–93, 2020.
[50] T. M. Alam, K. Shaukat, H. Mahboob, M. U. Sarwar, F. Iqbal, A. Nasir, I. A. Hameed, and S. Luo, "A machine learning approach for identification of malignant mesothelioma etiological factors in an imbalanced dataset," *Comput. J.*, May 2021.
[51] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. Int. Joint Conf. Artif. Intell.*, vol. 17, no. 1. Seattle, WA, USA: Lawrence Erlbaum Associates, 2001, pp. 973–978.
[52] B. Liu, Y. Ma, and C. K. Wong, "Improving an association rule based classifier," in *Proc. Eur. Conf. Princ. Data Mining Knowl. Discovery*. Berlin, Germany: Springer, 2000, pp. 504–509.
[53] A. Majid, S. Ali, M. Iqbal, and N. Kausar, "Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines," *Comput. Methods Programs Biomed.*, vol. 113, no. 3, pp. 792–808, Mar. 2014.
[54] Q. Wen, L. Sun, X. Song, J. Gao, X. Wang, and H. Xu, "Time series data augmentation for deep learning: A survey," 2020, *arXiv:2002.12478*. [Online]. Available: https://arxiv.org/abs/2002.12478
[55] S. Ahmed, A. Mahbub, F. Rayhan, R. Jani, S. Shatabda, and D. M. Farid, "Hybrid methods for class imbalance learning employing bagging with sampling techniques," in *Proc. 2nd Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solution (CSITSS)*, Dec. 2017, pp. 1–5.
[56] S. M. A. Elrahman and A. Abraham, "Class imbalance problem using a hybrid ensemble approach," *Int. J. Hybrid Intell. Syst.*, vol. 12, no. 4, pp. 219–227, Mar. 2016.
[57] T. M. Alam, K. Shaukat, I. A. Hameed, S. Luo, M. U. Sarwar, S. Shabbir, J. Li, and M. Khushi, "An investigation of credit card default prediction in the imbalanced datasets," *IEEE Access*, vol. 8, pp. 201173–201198, 2020, doi: 10.1109/ACCESS.2020.3033784.
[58] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data," Univ. California, Berkeley, Berkeley, CA, USA, Tech. Rep. 666, 2004, p. 24, vol. 110, nos. 1–12.
[59] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 2, pp. 539–550, Apr. 2008.
[60] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: Improving classification performance when training data is skewed," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.
[61] K. Shaukat, S. Zaheer, and I. Nawaz, "Association rule mining: An application perspective," *Int. J. Comput. Sci. Innov.*, vol. 2015, no. 1, pp. 29–38, 2015.
[62] F. Jiang, X. Yu, H. Zhao, D. Gong, and J. Du, "Ensemble learning based on random super-reduct and resampling," *Artif. Intell. Rev.*, vol. 54, no. 4, pp. 3115–3140, Apr. 2021.
[63] S. Tewari, U. D. Dwivedi, and S. Biswas, "A novel application of ensemble methods with data resampling techniques for drill bit selection in the oil and gas industry," *Energies*, vol. 14, no. 2, p. 432, Jan. 2021.
[64] Y. Qian, Y. Liang, M. Li, G. Feng, and X. Shi, "A resampling ensemble algorithm for classification of imbalance problems," *Neurocomputing*, vol. 143, pp. 57–67, Nov. 2014.
[65] M. Lázaro, F. Herrera, and A. R. Figueiras-Vidal, "Ensembles of cost-diverse Bayesian neural learners for imbalanced binary classification," *Inf. Sci.*, vol. 520, pp. 31–45, May 2020.
[66] X. Yang, M. Khushi, and K. Shaukat, "Biomarker CA125 feature engineering and class imbalance learning improves ovarian cancer prediction," in *Proc. IEEE Asia–Pacific Conf. Comput. Sci. Data Eng. (CSDE)*, Dec. 2020, pp. 1–6.
[67] H. Barlow, S. Mao, and M. Khushi, "Predicting high-risk prostate cancer using machine learning methods," *Data*, vol. 4, no. 3, p. 129, Sep. 2019.
[68] M. Mukherjee and M. Khushi, "SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features," *Appl. Syst. Innov.*, vol. 4, no. 1, p. 18, Mar. 2021.
[69] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Inf. Sci.*, vol. 513, pp. 429–441, Mar. 2020.
[70] Q. Gu, L. Zhu, and Z. Cai, "Evaluation measures of the classification performance of imbalanced data sets," in *Proc. Int. Symp. Intell. Comput. Appl.* Berlin, Germany: Springer, 2009, pp. 461–471.
[71] E. Mahmoudi, N. Kamdar, N. Kim, G. Gonzales, K. Singh, and A. K. J. B. Waljee, "Use of electronic medical records in development and validation of risk prediction models of hospital readmission: Systematic review," *BMJ*, vol. 369, pp. 1–10, Apr. 2020.
[72] 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, Apr. 2013.
[73] Y. Feng, M. Zhou, and X. Tong, "Imbalanced classification: An objective-oriented review," 2020. *arXiv:2002.04592*. [Online]. Available: https://arxiv.org/abs/2002.04592
[74] N. V. Chawla, "Data mining for imbalanced datasets: An overview," in *Data Mining and Knowledge Discovery Handbook*. Boston, MA, USA, 2009, pp. 875–886.
[75] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, 2016.
[76] K. Shaukat, T. M. Alam, S. Luo, S. Shabbir, I. A. Hameed, J. Li, S. K. Abbas, and U. Javed, "A review of time-series anomaly detection techniques: A step to future perspectives," Springer, Cham, Switzerland, Tech. Rep. 978-3-030-73100-7_60, 2021.
[77] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, 1997.
[78] K. Shaukat, S. Luo, S. Chen, and D. Liu, "Cyber threat detection using machine learning techniques: A performance evaluation perspective," in *Proc. Int. Conf. Cyber Warfare Secur. (ICCWS)*, Oct. 2020, pp. 1–6.
[79] T. M. Alam and M. J. Awan, "Domain analysis of information extraction techniques," *Int. J. Multidisciplinary Sci. Eng.*, vol. 9, no. 6, pp. 1–9, 2018.
[80] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, Mar. 2005.
[81] N. Mahfouz, I. Ferreira, S. Beisken, A. von Haeseler, and A. E. Posch, "Large-scale assessment of antimicrobial resistance marker databases for genetic phenotype prediction: A systematic review," *J. Antimicrobial Chemotherapy*, vol. 75, no. 11, pp. 3099–3108, Nov. 2020.
[82] T. Meraj, A. Hassan, S. Zahoor, H. T. Rauf, M. I. Lali, L. Ali, and S. A. C. Bukhari, "Lungs nodule detection using semantic segmentation and classification with optimal features," Neural Comput. Appl., U.K., Tech. Rep. s00521-020-04870-2, 2019, pp. 1–14.
[83] A. Baldominos, A. Puello, H. Ogul, T. Asuroglu, and R. Colomo-Palacios, "Predicting infections using computational intelligence—A systematic review," *IEEE Access*, vol. 8, pp. 31083–31102, 2020, doi: 10.1109/ACCESS.2020.2973006.
[84] S. Albahli, H. T. Rauf, A. Algosaibi, and V. E. Balas, "AI-driven deep CNN approach for multi-label pathology classification using chest X-rays," *PeerJ Comput. Sci.*, vol. 7, p. e495, Apr. 2021.
[85] J. Latif, C. B. Xiao, S. S. Tu, S. U. Rehman, A. Imran, and A. Bilal, "Implementation and use of disease diagnosis systems for electronic medical records based on machine learning: A complete review," *IEEE Access*, vol. 8, pp. 150489–150513, 2020.
[86] S. Albahli, H. T. Rauf, M. Arif, M. T. Nafis, and A. Algosaibi, "Identification of thoracic diseases by exploiting deep neural networks," *Comput., Mater. Continua*, vol. 66, no. 3, pp. 3139–3149, 2021.
[87] H. T. Rauf, M. I. U. Lali, M. A. Khan, S. Kadry, H. Alolaiyan, A. Razaq, and R. Irfan, "Time series forecasting of COVID-19 transmission in Asia Pacific countries using deep neural networks," *Pers Ubiquitous Comput.*, pp. 1–18, Jan. 2021, doi: 10.1007/s00779-020-01494-0.

[88] M. M. Islam, F. Karray, R. Alhajj, and J. Zeng, "A review on deep learning techniques for the diagnosis of novel coronavirus (COVID-19)," *IEEE Access*, vol. 9, pp. 30551–30572, 2021, doi: 10.1109/ACCESS.2021.3058537.
[89] C. S. Zhu, P. F. Pinsky, B. S. Kramer, P. C. Prorok, M. P. Purdue, C. D. Berg, and J. K. Gohagan, "The prostate, lung, colorectal, and ovarian cancer screening trial and its associated research resource," *J. Nat. Cancer Inst.*, vol. 105, no. 22, pp. 1684–1693, Nov. 2013.
[90] C. M. Tammemagi, P. F. Pinsky, N. E. Caporaso, P. A. Kvale, W. G. Hocking, T. R. Church, T. L. Riley, J. Commins, M. M. Oken, C. D. Berg, and P. C. Prorok, "Lung cancer risk prediction: Prostate, lung, colorectal and ovarian cancer screening trial models and validation," *J. Nat. Cancer Inst.*, vol. 103, no. 13, pp. 1058–1068, Jul. 2011.
[91] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[92] S. Kamran, I. Farhat, A. T. Mahboob, A. G. Kaur, D. Liton, K. A. Ghaffar, I. Rimsha, S. Irum, and R. Afifah, "The impact of artificial intelligence and robotics on the future employment opportunities," *Trends Comput. Sci. Inf. Technol.*, vol. 5, no. 1, pp. 50–54, Sep. 2020.
[93] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern Recognit.*, vol. 48, no. 9, pp. 2839–2846, 2015.
[94] E. Bisong, "Introduction to scikit-learn," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Berkeley, CA, USA: Springer, 2019, pp. 215–229.
[95] M. Z. Latif, K. Shaukat, S. Luo, I. A. Hameed, F. Iqbal, and T. M. Alam, "Risk factors identification of malignant mesothelioma: A data mining based approach," in *Proc. Int. Conf. Electr., Commun., Comput. Eng. (ICECCE)*, Jun. 2020, pp. 1–6.
[96] N. Noorhalim, A. Ali, and S. M. Shamsuddin, "Handling imbalanced ratio for class imbalance problem using SMOTE," in *Proc. 3rd Int. Conf. Comput., Math. Statist. (iCMS)*. Singapore: Springer, 2019, pp. 19–30.
[97] Y. Ali, A. Farooq, T. M. Alam, M. S. Farooq, M. J. Awan, and T. I. Baig, "Detection of schistosomiasis factors using association rule mining," *IEEE Access*, vol. 7, pp. 186108–186114, 2019.
[98] M. U. Ghani, T. M. Alam, and F. H. Jaskani, "Comparison of classification models for early prediction of breast cancer," in *Proc. Int. Conf. Innov. Comput. (ICIC)*, Nov. 2019, pp. 1–6.
[99] T. M. Alam, K. Shaukat, M. Mushtaq, Y. Ali, M. Khushi, S. Luo, and A. Wahab, "Corporate bankruptcy prediction: An approach towards better corporate world," *Comput. J.*, Jun. 2020.
[100] T. M. Alam, M. Milhan, M. Atif, A. Wahab, and M. Mushtaq, "Cervical cancer prediction through different screening methods using data mining," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 2, pp. 388–396, 2019.

MATLOOB KHUSHI received the Ph.D. degree from The University of Sydney, in 2016. He holds academic positions at The University of Sydney and the University of Suffolk, U.K. His research interests include contributing novel algorithms for life sciences and financial trading.

KAMRAN SHAUKAT received the M.Sc. degree in computer science from Mohammad Ali Jinnah University, Pakistan. He is currently pursuing the Ph.D. degree with The University of Newcastle, Callaghan, NSW, Australia. He has served the University of the Punjab, Pakistan, for seven years, as a Lecturer. He is the author of many articles in machine learning, databases, and cyber security. He has served as a Reviewer to many journals, including IEEE Access. He has attended several international conferences, including the USA, U.K., Thailand, Turkey, and Pakistan. He received the Gold Medal for his M.Sc. degree.

TALHA MAHBOOB ALAM received the bachelor's degree in software engineering from the University of Management and Technology (UMT), Lahore, Pakistan, in 2017, and the master's degree in computer science from the University of Engineering and Technology (UET), Lahore, in 2020. He is currently serving for the Virtual University of Pakistan. He has published more than 25 journals and conference papers of international repute. His research interests include big data, machine learning, deep learning, and knowledge discovery in databases.



IBRAHIM A. HAMEED is currently a Professor and the Deputy Head of the research and innovation, and the Head of the international master program in simulation and visualization with the Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, Trondheim, Norway, where he has been an Associate Professor with the Department of ICT and Natural Sciences, since 2015.

SHAHADAT UDDIN received the Ph.D. degree in complex networks and health analytics from The University of Sydney, in 2011. He is currently a Senior Lecturer with the Faculty of Engineering, School of Project Management, The University of Sydney.

SUHUAI LUO received the bachelor's and master's degrees from Nanjing University of Posts and Telecommunications, and the Ph.D. degree from The University of Sydney, all in electrical engineering. He is currently an Associate Professor in information technology with The University of Newcastle. His main research interests include image processing, computer vision, machine learning, cybersecurity, and media data mining. His diverse research focus has led him to conduct studies in areas ranging from medical imaging for computer-aided diagnoses to computer vision for intelligent driving systems and machine learning for enhancing cybersecurity.

XIAOYAN YANG (Member, IEEE) received the Bachelor of Engineering degree in mechatronic engineering from Beijing Jiaotong University, in 2019, the Bachelor of Engineering degree (Hons.) from the University of Wollongong, Australia, and the master's degree in data science from The University of Sydney, Australia, in 2021.

MARANATHA CONSUELO REYES received the B.S. degree in mathematics and the M.S. degree in applied mathematics from the University of the Philippines, and the Master of Data Science degree from The University of Sydney, Australia, in 2020.