Informatics in Medicine Unlocked 25 (2021) 100690

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Performance analysis of cost-sensitive learning methods with application to imbalanced medical data

Ibomoiye Domor Mienye, Yanxia Sun *

Department of Electrical and Electronic Engineering Science, University of Johannesburg, Johannesburg, 2006, South Africa

ARTICLE INFO

Keywords:
Cost-sensitive learning
Imbalanced classification
Machine learning
Medical diagnosis

ABSTRACT

Many real-world machine learning applications require building models using highly imbalanced datasets. Usually, in medical datasets, the healthy patients or samples are dominant, making them the majority class, while the sick patients are few, making them the minority class. Researchers have proposed numerous machine learning methods to predict medical diagnosis. Still, the class imbalance problem makes it difficult for classifiers to adequately learn and distinguish between the minority and majority classes. Cost-sensitive learning and resampling techniques are used to deal with the class imbalance problem. This research focuses on developing robust cost-sensitive classifiers by modifying the objective functions of some well-known algorithms, such as logistic regression, decision tree, extreme gradient boosting, and random forest, which are then used to efficiently predict medical diagnosis. Meanwhile, as opposed to resampling techniques, our approach does not alter the original data distribution. Firstly, we implement the standard versions of these algorithms to provide a baseline for performance comparison. Secondly, we develop their corresponding cost-sensitive algorithms. For the proposed approaches, it is not necessary to change the distribution of the original data as the modified algorithms consider the imbalanced class distribution during training, thereby resulting in more reliable performance than when the data is resampled. Four popular medical datasets, including the Pima Indians Diabetes, Haberman Breast Cancer, Cervical Cancer Risk Factors, and Chronic Kidney Disease datasets, are used in the experiments to validate the performance of the proposed approach. The experimental results show that the cost-sensitive methods yield superior performance compared to the standard algorithms.

1. Introduction

The advances in healthcare technology and machine learning (ML) have saved several lives through efficient disease prediction, patient monitoring, and clinical decision making [1]. These advances have also made available numerous medical data. There is a need for further research and development to avoid the inaccurate prediction of diseases, which can be dangerous for the patients [2]. Meanwhile, in ML research, one problem that has been widely studied is the class imbalance problem. By definition, the class imbalance can be referred to as a phenomenon where the majority class exceeds that of the minority class by a huge factor [3,4]. Research has shown that medical data are mostly imbalanced [5], where the majority class (negative or healthy patients) significantly outnumber the minority class (positive or sick patients). Usually, most ML algorithms used for binary classifications tasks assume an even distribution of the classes. Hence, when trained with imbalanced data, the model gets dominated by samples from the majority class, thereby degrading the model's performance [6]. This problem is so crucial that it is viewed as one of the ten big challenges in machine learning research [7].

Furthermore, ML algorithms assume that misclassification errors (false negative and false positive) are equal [8]. However, this assumption can be dangerous in imbalanced classification problems such as medical diagnosis, fraud detection, and access control systems [9]. For example, misclassifying a positive instance is more costly than misclassifying a negative sample. Meanwhile, resampling techniques have been used to balance the class distributions in imbalanced datasets [10]. Resampling methods aim to manually balance the data through undersampling the majority instances or oversampling the minority instances; sometimes, both methods are used. However, resampling techniques may omit some possible valuable data and increase the computational cost with unnecessary instances. In essence, both undersampling and oversampling methods changes the distribution of the various classes [11].

* Corresponding author.
E-mail address: ysun@uj.ac.za (Y. Sun).

https://doi.org/10.1016/j.imu.2021.100690
Received 4 June 2021; Received in revised form 30 July 2021; Accepted 1 August 2021
Available online 3 August 2021

Meanwhile, another method exists called cost-sensitive learning (CSL) that considers the cost associated with the misclassification of samples [12]. Rather than artificially creating balanced class distributions via sampling techniques, cost-sensitive learning solves the imbalanced class problem by utilizing cost matrices that outline the costs associated with the misclassification of the various classes [13]. By definition, cost-sensitive learning can be considered a subfield of ML that considers the cost of classification errors during model training [8]. Research has shown that cost-sensitive learning yields enhanced performance in applications where the dataset has a skewed class distribution [14].

Generally, ML algorithms aim to minimize error during training, and several functions can be utilized to compute the error or loss of a model on training data. In cost-sensitive learning, a penalty is placed for misclassifications, and this is referred to as the cost. Cost-sensitive learning aims to minimize the misclassification cost of a model on the input data. Hence, instead of optimizing the accuracy, the algorithm tries to minimize the total misclassification cost [15]. Furthermore, recent research has suggested a high correlation between cost-sensitive learning and imbalanced classification; hence, the conceptual frameworks and algorithms utilized for cost-sensitive learning can be inherently employed for imbalanced classification tasks [16]. Also, some research works have demonstrated that when attempting to solve imbalanced classification problems, cost-sensitive learning leads to superior performance [11], and it is a more suitable approach than sampling techniques.

Several research works have proposed numerous methods to classify imbalanced medical data, as stated in Refs. [5,14]. However, most of these methods focus on data resampling, such as in Refs. [3,17,18]. Even though there has been numerously published papers regarding the classification of imbalanced medical data, the focus has been on resampling methods. This research aims to provide a general overview of the imbalanced classification problem and ML algorithms suitable for such classification problems focusing on medical data. In the process, we develop some cost-sensitive ML algorithms to conduct a comparative study with standard algorithms. Also, while other research works on cost-sensitive learning have proposed single CSL algorithms, this research work implements numerous CSL algorithms and analyze the prediction performance between the standard and cost-sensitive ML algorithms on selected medical diagnosis datasets. The algorithms studied are logistic regression (LR), decision tree (DT), extreme gradient boosting (XGBoost), and random forest (RF). Meanwhile, we employ four medical datasets in this research, including the Pima Indians Diabetes, Haberman Breast Cancer, Cervical Cancer Risk Factors, and Chronic Kidney Disease datasets.

The rest of this paper is structured as follows. In Section 2, we briefly review some related works. Section 3 introduces the cost-sensitive learning framework and the algorithms utilized in this work. Section 4 discusses the datasets and performance assessment criteria used in this paper. In Section 5, we present the experimental results, followed by analyzing and discussing these results. Lastly, Section 6 concludes the research and discuss some future research directions.

2. Literature review

This section provides a brief overview and review of some related works regarding the class imbalance problem and cost-sensitive learning approaches in medical diagnosis.

2.1. Overview of the class imbalance problem

Despite the recent advancements in machine learning and deep learning, the class imbalance problem remains a challenge for researchers [19]. In binary classification tasks with data examples from two classes or groups, the data is said to have a class imbalance when one group, the minority class, have lesser instances than the other group, the majority class. In numerous imbalanced classification problems, the class of interest is the minority class, i.e., the positive or sick patients in medical data. In medical diagnosis applications, most patients do not have the disease (i.e., majority or negative class) and predicting those with the disease is of paramount importance. Therefore, it is challenging to learn and make accurate predictions using imbalanced medical data, and non-traditional ML algorithms are usually required to obtain suitable performance. Also, in imbalanced medical data, ML models usually overclassify the majority class because of their higher prior probability. Hence, the samples in the minority class are misclassified more than those in the majority class [19].

Numerous studies have proposed methods to solve the imbalanced classification problem. For example, Kuo et al. [20] proposed a technique using the information granulation (IG) concept. The proposed algorithm balances the class ratio in the data by gathering the samples from the majority class into granules. The first step in the algorithm uses metaheuristic techniques such as genetic algorithm K-means, particle swarm optimization, and artificial bee colony K-means to generate a set of IGs. The second step uses a classifier to predict prostate cancer survival rate using patient data. Similarly, Liu et al. [21] proposed a two-step ML technique to predict cerebral stroke using an imbalanced dataset. The first step employs random forest regression for missing data imputation, and the second step uses an automated hyperparameter optimization-based deep neural network to classify the imbalanced data. The approach achieved an enhanced cerebral stroke prediction.

Several prior works to solve the class imbalance problem can be split into data-level and algorithmic-level approaches. Data-level approaches modify the class distribution of the data via resampling methods to create a balanced dataset. Though resampling techniques have been widely utilized, they have some shortcomings since they alter the original class distribution of the data [22]. Precisely, undersampling can remove important information that may be vital in the learning process, while oversampling can result in overfitting and sometimes increase the computational cost [22]. Blagus and Lusa [23] presented a detailed theoretical and empirical study of resampling, focusing on Synthetic Minority Oversampling Technique (SMOTE). The study applied SMOTE to several real and simulated imbalanced datasets to explain the behaviour of the algorithm.

Meanwhile, the SMOTE algorithm has been widely applied in medical diagnosis to provide a class balance in imbalanced datasets. For example, Zeng et al. [24] combined SMOTE with the Tomek links technique to balance three medical datasets, which improved the eight classifiers' performance in the study. Also, Xu et al. [25] proposed an improved method to classify imbalanced medical data by combining misclassification-oriented SMOTE (M-SMOTE) and edited nearest neighbor (ENN), while a random forest classifier is used to classify the samples. The study utilized ten imbalanced medical datasets, and the proposed method obtained improved performance compared to other classical resampling techniques. Shilaskar et al. [26] used SMOTE and modified particle swarm optimization (M-PSO) method to balance a medical dataset. The study employed five ML classifiers to classify the resampled data and analyze the resampling techniques' robustness.

In contrast to data-level approaches, algorithm-level methods alter the classifier, for example, ensemble learning and cost-sensitive learning. Ensemble learning methods utilize multiple learning algorithms to achieve better classification performance than when the individual algorithms are used [5]. Some recently proposed ensemble learning methods for medical diagnosis [27,28] have achieved good performance. Zhu et al. [29] proposed a method to classify high-dimensional imbalanced medical data using an algorithm that combines random forest and feature selection techniques. The technique involves dimensionality reduction of the high dimensional data and classification of the target variables. The experimental results show that the approach achieved good classification accuracy when applied to high dimensional medical data.

Furthermore, a few hybrid ensemble methods [5,17] have been proposed; these methods combine resampling and ensemble learning

techniques. An in-depth review of ensemble learning methods for imbalanced classification is presented in Ref. [30]. Although ensemble learning leads to improved performance, the combination of multiple classifiers is a complex process and result in higher training times. Recent research [11] has shown that cost-sensitive learning can ensure the algorithm correctly classifies the minority class and does not impact its complexity or computational time.

2.2. Research on cost-sensitive learning for medical diagnosis

Many research works on the application of machine learning to medical diagnosis usually employ traditional ML algorithms and improved algorithms via ensemble learning [27,29], artificial neural networks (ANN) [31], evolutionary algorithms [32], sparse autoencoders (SAE) [33], among others. However, a few research works have employed cost-sensitive learning to medical diagnosis. Cost-sensitive learning involves modifying the algorithm's objective function to ensure it focuses more on accurately predicting the minority class. Recently, cost-sensitive learning was applied to classify chronic kidney disease (CKD) in Ref. [34]; the research work proposed a cost-sensitive ensemble method that incorporates feature ranking capabilities. The proposed approach was compared with seven classification algorithms and eight feature selection techniques to demonstrate the robust performance of the CSL approach, which performed better than the other algorithms. The research concluded that cost-sensitive learning is an accurate and cost-effective approach to solving CKD's imbalanced classification.

Cost-sensitive learning has also been utilized to detect breast cancer, one of the most prevalent cancers among women. Breast cancer classification is difficult to achieve due to the skewed class distribution of the dataset, thereby leading to poor performance when standard ML algorithms are applied for this classification. In Ref. [35], a cost-sensitive XGBoost was developed with application to breast cancer detection, and the study utilized four breast cancer datasets with uneven class distribution. The results showed that the cost-sensitive XGBoost achieved excellent performance in all four datasets. Furthermore, a cost-sensitive decision tree classifier was developed by integrating game theory [36]. The algorithm used the concept of lever pulls from the multi-armed bandit game in choosing the features during tree formulation via a feed-forward framework to obtain features that maximize the reward. The proposed method was experimented on 15 datasets, including datasets to classify breast cancer, diabetes, heart disease, hepatitis etc. The experimental result showed that the proposed method obtained superior performance.

Furthermore, Zieba et al. [37] proposed an adaptive boosting based support vector machine (SVM) to handle the imbalanced classification of lung cancer patients post-operation life expectancy. The method combines the advantage of ensemble learning and cost-sensitive SVM. The proposed cost-sensitive classifier obtained enhanced performance when compared to other popular classifiers used to handle imbalanced data. Similarly, Ali et al. [38] developed a method that combines cost-sensitive learning and ensemble learning techniques to predict breast cancer. The ensemble learning methods considered in the study include GentleBoost, Bagging, and adaptive boosting. The experimental results show that the cost-sensitive GentleBoost performed better than other ensemble classifiers.

In another study, Wan et al. [39] presented a novel cost-sensitive learning-based boosting algorithm called RankCost to predict imbalanced medical data. The method uses a ranking function to maximize the difference between the majority and minority classes. The ranking function assigns higher scores to instances in the minority class than instances in the majority class. Zhu et al. [40] developed a cost-sensitive random forest to deal with the imbalanced class problem in medical diagnosis. The study employed several medical datasets, and the proposed algorithm showed improved performance, specifically in accurately predicting both the minority and majority classes.

Furthermore, Gan et al. [41] incorporated a tree-augmented naïve Bayes algorithm and cost-sensitive adaptive boosting (AdaCost) algorithm with application to imbalanced medical data. The proposed algorithm was tested on several medical data, including the cervical cancer risk factors dataset and Cleveland heart disease dataset. The experimental results show that the proposed algorithm performed better than some state-of-the-art methods.

Cost-sensitive neural networks have also been developed; in Ref. [42], a cost-sensitive deep learning approach was proposed to predict hospital readmission. The early prediction of hospital readmission ensures timely intervention of medical practitioners, which is necessary to prevent serious complications. The approach involves automatic feature learning of the patient data using convolutional neural networks (CNN) combined with a cost-sensitive multilayer perceptron (MLP) classifier. In addition, the cost-sensitive MLP ensured the class imbalance was considered during model training. Finally, the approach was applied to a real-world medical dataset. It achieved an area under the receiver operating characteristics curve (AUC) value of 0.70, which was superior to the baseline models.

Furthermore, Wu et al. [43] proposed a novel cost-sensitive radial basis function neural network (RBF-NN) for medical diagnosis. The method involves using a genetic algorithm and an enhanced PSO to optimize the parameters and structure of the cost-sensitive RBF-NN. When applied to five medical datasets, the experimental results show that the proposed method obtains better accuracy and AUC values than some state-of-the-art methods. Meanwhile, this paper aims to build on prior research by providing a detailed performance analysis of cost-sensitive learning algorithms with application to some medical datasets.

3. Materials and methods

In this section, we discuss the cost-sensitive learning approach and the various algorithms implemented in this research.

3.1. Cost-sensitive learning

For a binary classification problem, assuming $D = \{(x_i, y_i)\}_{i=1}^{n}$ represents a training set with n independent and identically distributed random variables, where $x_i \in X \subseteq \mathbb{R}^d$ is the ith instance and $y_i \in Y = \{-1, 1\}$ is the ith equivalent dependent variable. To achieve the goal of classification, a predictor $f : X \rightarrow \mathbb{R}$ is obtained, and a classification rule is often considered to be $sign[f(x)]$. To measure the performance, a nonnegative loss function $L : \mathbb{R} \times Y \rightarrow \mathbb{R}$ is used. Therefore, the regularised empirical risk minimization (ERM) is expressed thus:

$$\min_{f \in F} J(f, D) = \min_{f \in F} \left\{ \frac{1}{n} \sum_{i=1}^{n} L(f(x_i), y_i) + \lambda N(f) \right\} \quad (1)$$

Here, λ denotes the regularisation parameter, whereas $N(.)$ represents the regularizer to prevent overfitting [44]. Generally, most ML algorithms achieve error minimization using the ERM configuration. These algorithms assume that all the misclassification errors have the same cost, resulting in classifiers that are not cost-sensitive. In reality, many machine learning problems, such as medical diagnosis [44] and fraud detection [45,46], are cost-sensitive.

Cost-sensitive learning is a special type of learning where misclassification costs are taken into consideration. Cost-sensitive learning aims to minimize the total cost. It differs from cost-insensitive learning because it handles distinct misclassifications distinctively, i.e., the cost of classifying a sick patient as healthy is different from the cost of predicting a healthy patient as sick. In contrast, cost-insensitive learning aims to minimize the error rate and neglect the various misclassification errors. Furthermore, cost-insensitive classifiers assume that all the misclassification costs are equal. However, this assumption is not valid in most ML applications [47]. For example, in predicting diseases such

Table 1
The cost matrix.

	Actual negative ($y = -1$)	Actual positive ($y = 1$)
Predictive negative *sign* $[f(x)] = -1$	$C(-1, -1) = C_{TN}$	$C(-1, 1) = C_{FN}$
Predicted positive *sign* $[f(x)] = 1$	$C(1, -1) = C_{FP}$	$C(1, 1) = C_{TP}$

as cancer, a misclassification (i.e., false negative) is costlier than a false positive because the patient can die due to delayed treatment occasioned by the misclassification.

Cost-sensitive learning considers uneven misclassification costs. Most times, there is zero cost for correct classifications, that is, $C_{TN} = C_{TP} = 0$. Furthermore, classifying an instance incorrectly often have more cost than classifying it correctly (i.e., $C_{FN} > C_{TP}$ and $C_{FP} > C_{TN}$); the cost matrix is shown in Table 1. For cost-insensitive classifiers, $C_{FP} = C_{FN}$, and for cost-sensitive classifiers $C_{FP} \neq C_{FN}$. Also, for medical diagnosis, the cost of false negatives is usually more than the cost of false positives (i.e., $C_{FN} > C_{FP}$).

Meanwhile, it is possible to formulate a classification problem regarding risk minimization for a given cost matrix by modifying the loss function. By modelling the loss function to consider variable misclassification cost, we can obtain a cost-sensitive classifier. Also, there are numerous methods designed by weighting the ERM loss function of type $L(f(x),y) = L(yf(x))$, i.e., a margin-based loss function [48]. Margin-based loss functions are essential in binary classification because, unlike other loss functions, they do not consider the difference between the actual label and prediction. Rather, they penalize predictions based on how well they correlate with the sign of the target. For a function f and tuple (x,y), the margin of the tuple obtained by f can be represented as $yf(x)$ [48]. Therefore, a cost-sensitive classifier can then be formulated by minimizing the empirical risk:

$$J_c(f, D) = \frac{1}{n}\sum_{i=1}^{n} g(y_i)L(y_i(h(y_i)f(x_i) + \eta)) + \lambda N(f) \tag{2}$$

From (2), $g(y_i)$ denotes a sample-based weight function while $h(y_i)$ denotes a margin-based weight function, and η signifies the weight constant, and these parameters are connected to the target variable and represent the inequality in misclassification costs [49]. Therefore, using the necessary weighting approach, the summation in (2) can be considered to evaluate the cumulative misclassification cost of the classifier $f(x)$. Furthermore, diverse CSL methods can be proposed via the combination of different options of $L, h, g,$ and η. However, in this paper, we develop CSL classifiers based on an instance where $\eta = 0$, $h(y) = 1$, and

$$g(y) = \begin{cases} C_{FN} - C_{TP} \text{ if } y = 1 \\ C_{FP} - C_{TN} \text{ if } y = -1 \end{cases} \tag{3}$$

The subsequent sections will discuss how the selected algorithms are modified in line with this methodology to make them cost-sensitive.

3.2. Cost-sensitive logistic regression

Logistic regression in its standard form does not consider the imbalanced nature of some datasets; like most machine learning algorithms, it assumes an even class distribution. Therefore, it is crucial to modify the algorithm to consider the imbalanced class problem. To achieve this, a class weighting mechanism is employed to control how the algorithms' coefficients are updated during training. The weighting configuration ensures the model is penalized more for errors made on samples from the minority class. Also, the model is penalized less for errors made on samples from the majority class. Usually, the log-likelihood function $L(w)$ is expressed as:

$$L(w) = \sum_{i=1}^{n} [y_i \ln(P(y_i)) + (1 - y_i)\ln(1 - P(y_i))] \tag{4}$$

where $P(y_i)$ denotes the predicted probability that y is true for i [50]. Whereas the modified log-likelihood function can be represented as:

$$L(w) = \sum_{i=1}^{n} [C_{FP} y_i \ln(P(y_i)) + C_{FN}(1 - y_i)\ln(1 - P(y_i))] \tag{5}$$

This setting usually leads to a type of logistic regression that is well suited for imbalanced classification problems, and this is called a cost-sensitive logistic regression.

3.3. Cost-sensitive decision tree

Decision tree algorithms are efficient for classification problems when the class distribution in the dataset is balanced. However, they have poor performance when trained with imbalanced data. Usually, in decision trees, the split points are selected to optimally separate samples into two classes having minimal mixing (also called purity). However, if both sets have more samples from the majority class, then the instances from the minority class are abruptly being neglected. Meanwhile, to avoid this problem, we can modify the criteria utilized for split point selection to consider class importance, resulting in a cost-sensitive decision tree [36]. The purity is usually computed using the Gini index or entropy [4]. This paper implements an instance of classification and regression tree (CART); hence, the purity is computed using the Gini index [28]. The process of calculating the purity metric entails computing the probability of an instance being wrongly classified by the split. And the probability calculations include the summation of the number of instances in the various classes that make up a group.

Therefore, the criteria used for splitting can be updated to consider the purity of the split and be weighted by the importance of each class. We can achieve this by replacing the count of instances in the various groups with a weighted sum, where the coefficient is provided to weigh the sum. A large weight can then be given to the minority class, which is more important and has more influence on the node purity, and a lesser weight to the majority class, which has a lesser influence on the node purity. A general heuristic for the class weighting is to utilize the inverse of the class distribution of the dataset, i.e., for a class distribution of 10:100 ratio for the minority group to the majority group, the inverse would be to use 100 for the minority group and 10 for the majority group.

3.4. Cost-sensitive XGBoost

The XGBoost is an algorithm that uses the gradient boosting framework at its core. The XGBoost algorithm is an ensemble of decision trees, and it has been applied in diverse classification and regression tasks. It also has good performance in classification problems with uneven class distribution [51]. However, we can enhance the performance further by training the algorithm to focus more on the misclassification of the minority class, and the new algorithm is called a cost-sensitive XGBoost. Fortunately, for the XGBoost, the modification can be achieved by tuning a hyperparameter called scale_pos_weight in scikit-learn. In the XGBoost implementation, the default value for scale_pos_weight is 1.0. A good value for this hyperparameter would also be the inverse of the class distribution. We can use this hyperparameter to scale the errors made by the algorithm on the minority class during training, thereby prompting the algorithm to correct these errors. The model can therefore obtain improved performance when classifying instances in the minority class.

3.5. Cost-sensitive random forest

Random forest is an ensemble learning algorithm that is used for classification and regression. The algorithm constructs a multitude (or

Table 2
Summary of the datasets.

Dataset	Total samples	Number of samples in the majority class	Number of samples in the minority class
PID	767	500 (65.19 %)	267 (34.81 %)
Breast Cancer	305	224 (73.44 %)	81 (26.56 %)
Cervical Cancer	858	803 (93.59 %)	55 (6.41 %)
CKD	400	250 (62.5 %)	150 (37.5 %)

forest) of decision trees during training [52]. For classification tasks, the algorithm output the class that is the mode of the classes, and for regression, it outputs the mean prediction of the different decision trees. Furthermore, this algorithm rectifies the overfitting problem associated with decision trees. While the random forest is suitable for numerous applications, it has poor performance on imbalance classification tasks. Also, the data characteristics impact the performance of the random forest algorithm [28]. To modify the standard random forest to be cost-sensitive, we assign weights to the various classes. Also, we use the inverse of the class distribution, thereby forcing the algorithm to focus more on the minority class.

4. Datasets and assessment criteria

Four imbalanced medical datasets are used in this research, including the Pima Indians Diabetes (PID) [53], Haberman Breast Cancer [54], Cervical Cancer Risk Factors [55], and the Chronic Kidney Disease (CKD) [56] datasets obtained from the University of California, Irvine machine learning repository. Firstly, the PID dataset was prepared by the United States National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) after a study on female patients over 21 years old of the Pima Indians tribe in Arizona. The dataset aims to predict whether a patient has diabetes or not, using some diagnostic data such as insulin level, age, body mass index, number of pregnancies etc. Secondly, Haberman's breast cancer aims to predict whether a patient would live for a minimum of five years or not after undergoing breast cancer surgery. The dataset originated from the University of Chicago Hospital during a study on patients who underwent breast cancer surgeries. The features of this dataset are age, the number of positive nodes detected, year, and survival status.

Meanwhile, the cervical cancer dataset predicts if a woman would get cervical cancer based on certain risk factors, including medical history, lifestyle factors, and demographic details. We used the reduced feature set obtained by Fernandes et al. [57] for the cervical cancer data, which has been widely used in numerous cervical cancer studies. Lastly, the CKD dataset contains patient data such as blood pressure, red blood cells, serum creatinine, haemoglobin, anaemia, hypertension etc. Apollo Hospitals prepared the dataset in Tamilnadu, India. The dataset contains two classes, i.e. CKD and non-CKD, which corresponds to patients with chronic kidney disease and those without chronic kidney disease. Table 2 describes the distribution of samples in the various datasets. .

The breast cancer dataset does not contain missing values, while the cervical cancer and CKD dataset contain missing values. Meanwhile, the PID dataset does not explicitly contain missing values, but some biological measurements have a value of 0. Such incorrect measurements can negatively impact ML algorithms. Hence, we used the nearest neighbor imputation to predict and replace the missing values. The nearest neighbor imputation, which is an implementation of the k-nearest neighbor algorithm, is an effective method to estimate missing values [58]. The algorithm imputes a new value by taking the nearby samples in the dataset and computing their average. The scikit-learn ML library contains the KNNImputer class that is used to achieve nearest neighbor imputation. The number of neighbors set by the 'n-neighbors' hyperparameter is fixed to be 5, and the distance measure is the Euclidean distance, set by the 'metric' hyperparameter. Furthermore, the preprocessing carried out on the CKD dataset includes encoding categorical attributes into numerical values and feature scaling using MinMax Scaler.

Table 3
Confusion matrix.

	Actual positive	Actual negative
Predicted positive	True positive (*TP*)	False positive (*FP*)
Predicted negative	False negative (*FN*)	True negative (*TN*)

To evaluate the performance of the various algorithms developed in this paper, we utilize some assessment metrics such as accuracy, precision, recall, F-measure, and Cohen's kappa coefficient. These assessment metrics can be derived from the confusion matrix:

From the confusion matrix *TP*, *TN*, *FP*, and *FN* represent true positive, true negative, false positive, and false negative, respectively. True positive and true negative is the number of correct positive and negative predictions, respectively [59]. False positive is an error where the model incorrectly predicts a healthy patient as sick. In contrast, a false negative is an error where the model fails to predict the presence of a disease when it is present. The confusion matrix (see Table 3) provides a summary of the binary classification experimental results. The mathematical representation of the assessment metrics are detailed below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F\ measure = \frac{2 \times precision \times recall}{precision + recall} \tag{9}$$

$$Kappa = \frac{p_0 - p_c}{1 - p_c} \tag{10}$$

where p_0 represents the relative observed agreement among classifiers, whereas p_c denotes the probability that agreement is due to chance [60,61]. The Cohen's kappa coefficient (Kappa or κ) is a statistical test that was initially used to measure inter-rater reliability. Nowadays, Kappa statistic is utilized in ML mostly as a classifier performance measure because it compares the accuracy of a classifier to the accuracy of a random classifier. It was first introduced by Jacob Cohen [62] and has been widely used for binary and multiclass classification problems [63]. Kappa can also be obtained from the 2×2 confusion matrix used in ML and statistics to assess the performance of binary classifications:

$$Kappa = \frac{2 \times (TP \times TN - FP \times FN)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)} \tag{11}$$

The minimum κ value is -1, i.e., perfectly wrong predictions, and the maximum value is $+1$, i.e., perfect classifications. Meanwhile, when $\kappa \approx 0$, it implies the classifier's predictions are similar to random guessing [63].

Accuracy is the most commonly used metric when assessing the performance of binary classifiers [59,64]. However, it is not a suitable metric for imbalance classification problems [65] because it is influenced mainly by samples from the majority class. For instance, predicting all the samples as negative (majority class) in a highly imbalanced medical dataset would give a very high accuracy score. But in reality, the model has not learnt anything about the minority class. Hence, we consider other metrics that are more suitable for imbalanced classification tasks. Precision estimates the fraction of samples predicted to be positive that is truly positive. Recall (sensitivity or true positive rate) indicates the fraction of the positive examples correctly classified

Table 4
Assessment of the performance of the algorithms on the PID dataset.

Algorithm	Accuracy	Precision	Recall	F-Measure	Kappa
LR	0.751	0.689	0.710	0.699	0.739
CS LR	0.747	0.721	0.798	0.724	0.716
DT	0.703	0.630	0.711	0.668	0.665
CS DT	0.694	0.675	0.786	0.726	0.637
XGBoost	0.781	0.710	0.770	0.739	0.774
CS XGBoost	0.832	0.767	0.855	0.810	0.820
RF	0.758	0.725	0.710	0.717	0.716
CS RF	0.792	0.770	0.840	0.803	0.814

Table 5
Assessment of the performance of the algorithms on the breast cancer dataset.

Algorithm	Accuracy	Precision	Recall	F-Measure	Kappa
LR	0.742	0.758	0.701	0.729	0.682
CS LR	0.754	0.750	0.857	0.800	0.896
DT	0.644	0.699	0.684	0.691	0.471
CS DT	0.716	0.720	0.829	0.771	0.774
XGBoost	0.729	0.788	0.762	0.775	0.710
CS XGBoost	0.762	0.804	0.828	0.816	0.834
RF	0.707	0.747	0.754	0.751	0.716
CS RF	0.803	0.878	0.900	0.889	0.848

Table 6
Assessment of the performance of the algorithms on the cervical cancer dataset.

Algorithm	Accuracy	Precision	Recall	F-Measure	Kappa
LR	0.956	0.913	0.830	0.870	0.893
CS LR	0.940	0.942	0.978	0.960	0.914
DT	0.942	0.876	0.904	0.890	0.850
CS DT	0.933	0.918	0.920	0.919	0.831
XGBoost	0.981	0.978	0.961	0.969	0.969
CS XGBoost	0.986	1.000	1.000	1.000	0.982
RF	0.970	0.973	0.969	0.971	0.953
CS RF	0.988	1.000	1.000	1.000	0.989

Table 7
Assessment of the performance of the algorithms on the CKD dataset.

Algorithm	Accuracy	Precision	Recall	F-Measure	Kappa
LR	0.943	0.950	0.973	0.961	0.930
CS LR	0.979	0.974	1.000	0.987	0.964
DT	0.929	0.946	0.950	0.948	0.910
CS DT	0.951	0.925	1.000	0.961	0.938
XGBoost	0.940	0.957	0.921	0.939	0.959
CS XGBoost	0..981	0.973	1.000	0.986	0.974
RF	0.947	0.972	0.946	0.960	0.939
CS RF	0.986	0.990	1.000	0.995	0.983

as positive. The recall is an essential metric in imbalanced medical diagnosis because it solely depends on the minority class.

Meanwhile, precision and recall are usually combined to form a single metric called F-measure, another vital metric when dealing with datasets with a skewed class distribution. F-measure is the harmonic mean of precision and recall [64]. Soleymani et al. [59] investigated performance evaluation metrics used for imbalanced classification, focusing on F-measure, which is preferred over most metrics. The study further developed a novel F-measure global evaluation space, where a classifier's performance is represented by a curve that shows all the decision thresholds. The curves obtained by the F-measure space were then compared with precision-recall and ROC curves to demonstrate their suitability for imbalanced classification problems.

Ferri et al. [61] presented a detailed study and comparison of several ML performance metrics. The authors performed an experimental analysis of 18 performance evaluation metrics to study their behaviour. The study identified suitable performance metrics for diverse classification scenarios, including the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC). Therefore, other metrics used in this research to compare the performance of the various models are the ROC curve and the AUC. The AUC measures the model's ability to distinguish between the negative and positive classes [66]. Thus, a high AUC value demonstrates how good the model is at distinguishing the various classes.

5. Results and discission

In the cost-sensitive algorithms implemented in this paper, the misclassification cost relies on the sample class. denotes the misclassification cost of the negative class, while C_P denotes that of the positive class. Also, we utilize the assumption that the cost of correct classifications is zero ($C_{TN} = C_{TP} = 0$), therefore, $C_N = C_{FP}$ and $C_P = C_{FN}$. Furthermore, for the class weighting, a general heuristic was employed, i.e., using the inverse of the class distribution in the dataset. Therefore, the penalty for the wrong prediction of the minority class is more than the incorrect prediction of the majority class. This heuristic was chosen because it has led to improved results in previous works [8,10,49]. Furthermore, the cost-sensitive learning approach discussed in Section 3.1 is adapted to four algorithms, logistic regression, decision tree, XGBoost, and random forest. The experiments were conducted using a 16 GB RAM Windows computer with the following processor: Intel(R) Core(TM) i5-102100U CPU @ 1.60 GHz 2.10 GHz; the computations were carried out using the Python programming language and the scikit-learn ML library. The performance evaluation metrics discussed in Section 4 are utilized to measure the performance of the classifiers, and the repeated cross-validation, i.e., three repeats of 10-fold cross-validation, was used to evaluate the models.

5.1. Experimental results

The accuracy, precision, recall, F-measure, and Kappa assessment of the different classifiers are tabulated in Tables 4–7. Table 4 shows the performance when the classifiers are trained using the Pima Indians Diabetes dataset. Tables 5–7 show the performance when the classifiers are trained using the Haberman breast cancer, cervical cancer, and chronic kidney disease datasets. The first columns indicate the given classifier, while the various results are listed from the second to the last columns. The cost-sensitive version of the algorithms include cost-sensitive logistic regression (CS LR), cost-sensitive decision tree (CS DT), cost-sensitive XGBoost (CS XGBoost), and cost-sensitive random forest (CS RF).

From the experimental results, the cost-sensitive classifiers obtained superior performance compared to the cost-insensitive classifiers. The increased precision, recall, and F-measure values of the cost-sensitive models indicate an improved prediction of the minority class. For the Pima Indians Diabetes dataset, the cost-sensitive version of XGBoost obtained the best performance, followed by the cost-sensitive random forest. However, for the breast cancer, cervical cancer, and CKD datasets, the cost-sensitive random forest achieved the best performance, followed by the cost-sensitive XGBoost. Furthermore, it was observed that in all the datasets, the decision tree had the least performance for both the cost-sensitive and cost-insensitive models.

Meanwhile, some of the cost-sensitive models made more wrong predictions in the majority class than their cost-insensitive models, as observed from the accuracy values. Specifically, in Table 4, there was a decline in the accuracy values of the cost-sensitive versions of the logistic regression and decision tree. Also, in Table 6, the standard logistic regression and decision tree had a superior accuracy compared to their corresponding cost-sensitive models. This reduced accuracy values can be attributed to the wrong predictions in the majority class (false positive) [67]. From the experimental results, it can be seen that most of the cost-sensitive algorithms achieved κ values between the range

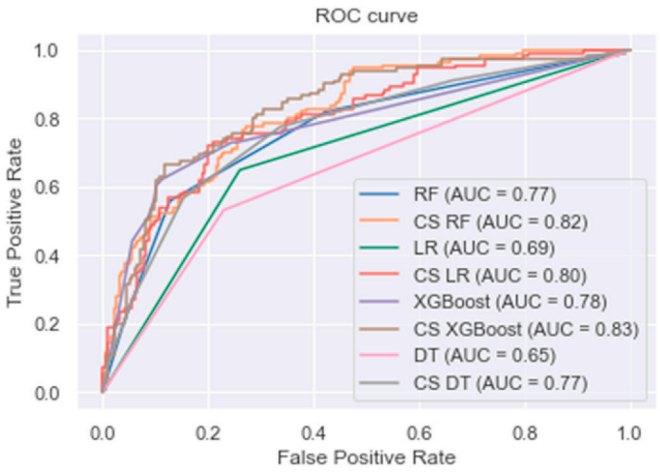

Fig. 1. ROC curves of the various classifiers trained with the PID dataset.

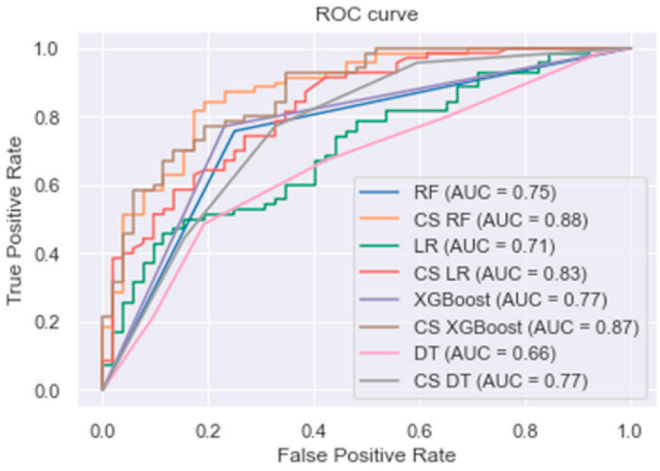

Fig. 2. ROC curves of the various classifiers trained with the breast cancer dataset.

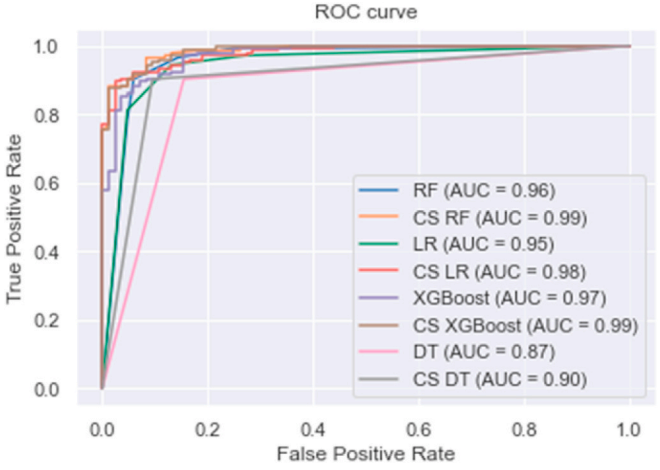

Fig. 3. ROC curves of the various classifiers trained with the cervical cancer dataset.

0.81–0.99, and according to McHugh et al. [68] the 0.81–0.99 range indicates near-perfect agreement between the raters. Also, only a few cost-sensitive classifiers obtained lower Kappa values than their corresponding cost-insensitive versions. These classifiers are CS LR and CS DT

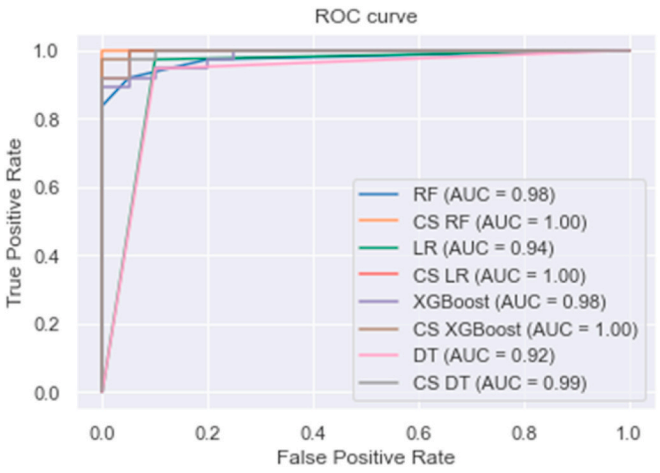

Fig. 4. ROC curves of the various classifiers trained with the CKD dataset.

Table 8
Comparison with other diabetes predictive models.

Reference	Method	AUC	Accuracy	Precision	Recall	F-Measure
Asniar et al. [69]	SMOTE + C4.5	0.792	0.751	0.716	0.829	0.768
Asniar et al. [69]	SMOTE + SVM	0.740	0.741	0.754	0.712	0.732
Chatrati et al. [70]	SVM	0.700	0.75	–	–	–
Hayashi and Yukita [71]	Recursive-rule extraction algorithm	–	0.8383	–	–	–
Khanam and Foo [72]	SVM + Feature selection	–	0.7682	0.761	0.768	0.759
Wei et al. [73]	Deep neural network	–	0.7784	–	–	–
Syed and Khan [74]	Decision forest	0.822	0.789	0.464	0.400	0.430
Abd El-Salam et al. [75]	Bayesian Nets	0.748	0.689	0.688	0.653	–
Pranto et al. [76]	SMOTE + RF	0.760	0.790	0.680	0.860	0.680
Pranto et al. [76]	SMOTE + KNN	0.710	0.730	0.570	0.660	0.610
This Paper	CS XGBoost	0.830	0.792	0.770	0.840	0.803

in Table 4 and CS DT in Table 6. Furthermore, Figs. 1–4 show the classifiers' ROC curves and the corresponding AUC values.

From the ROC curves and AUC values, it can be observed that the cost-sensitive models are more skilful in terms of predicting positive as positive and negative as negative, which further demonstrate the robustness of the cost-sensitive classifiers over the cost-insensitive classifiers. Finally, in Tables 8-11, the best performing cost-sensitive algorithms developed in this paper are used to compare with other research works, including resampling techniques, such as SMOTE and adaptive synthetic (ADASYN) sampling approach.

From Table 8 , it is observed that the cost-sensitive XGBoost obtained

Table 9
Comparison with other breast cancer predictive models.

Reference	Method	AUC	Accuracy	Precision	Recall	F-Measure
Gorunescu and Belciug [77]	MLP	–	0.761	–	–	–
Asniar et al. [69]	SMOTE + C4.5	0.730	0.705	0.726	0.655	0.689
Asniar et al. [69]	SMOTE + Naïve Bayes	0.671	0.620	0.773	0.336	0.469
Kaushik et al. [78]	Optimized SVM	–	0.788	–	–	–
Aljawad et al. [79]	SVM	–	0.744	0.747	0.737	–
Cahyana et al. [80]	ADASYN + Gradient Boosting	0.768	0.710	0.700	0.690	0.700
Cahyana et al. [80]	SMOTE + Gradient Boosting	0.763	0.670	0.670	0.640	0.650
Cahyana et al. [80]	Borderline SMOTE + Gradient Boosting	0.766	0.730	0.720	0.710	0.720
This paper	CS RF	0.880	0.803	0.878	0.900	0.889

Table 10
Comparison with other cervical cancer predictive models.

Reference	Method	AUC	Accuracy	Precision	Recall	F-Measure
Ijaz et al. [81]	SMOTE + RF	–	0.98925	0.98924	0.98936	0.98924
Ebiaredoh-Mienye et al. [82]	SAE + Softmax		0.970	0.980	0.950	0.970
Nithya and Ilango [83]	C5.0 + Feature selection	0.910	1.000	–	–	–
Wu and Zhou [84]	SVM	–	0.941	–	100	–
Abdoh et al. [85]	SMOTE + PCA + RF	–	0.957	–	0.977	–
Abdoh et al. [85]	SMOTE + RF	–	0.960	–	0.975	
Mienye et al. [33]	SAE + ANN	–	0.980	0.960	0.980	0.970
This paper	CS RF	0.990	0.988	1.000	1.000	1.000

Table 11
Comparison with other CKD predictive models.

Reference	Method	AUC	Accuracy	Precision	Recall	F-Measure
Khan et al. [64]	MLP	–	0.972	0.974	0.973	0.973
Rashed-Al-Mahfuz et al. [86]	Feature Selection + XGBoost	0.985	0.985	0.986	0.974	0.979
Ali et al. [34]	Ensemble learning + Feature selection	0.982	0.967	0.843	0.986	0.976
Ebiaredoh-Mienye et al. [82]	SAE + Softmax		0.980	0.970	0.970	0.970
Ogunleye and Wang [87]	Feature Selection + Optimized XGBoost	1.000	1.000	1.000	1.000	1.000
Chittora et al. [88]	SMOTE + ANN	0.996	0.964	0.981	0.913	0.946
Almustafa [89]	Feature Selection + Naïve Bayes	0.989	0.976	0.970	0.988	0.968
This paper	CS RF	1.000	0.986	0.990	1.000	0.995

comparable performance with other algorithms that used the Pima Indians Diabetes dataset. Furthermore, the cost-sensitive random forest also achieved excellent performance compared to other research works that used the breast cancer, cervical cancer, and CKD datasets, as seen in Tables 9-11 , respectively.

5.2. Discussion

From Section 5.1, it is observed that the cost-sensitive algorithms performed better than the standard algorithms in terms of precision, recall, F-measure, and AUC. A significant factor contributing to this improved performance is that giving more weight to the misclassifications of the minority class and penalizing the model more for wrong predictions of the minority class results in a model that pays more attention to this class. This forces the model to learn the instances in the minority class, which ultimately leads to a model that is skilful in predicting that class.

Secondly, the reduced accuracy in some cost-sensitive algorithms can be attributed to a few misclassifications in the majority class. These misclassifications impacted the accuracy since this metric indicates the ratio of correct predictions to the total predictions made. This is only normal as forcing the algorithm to focus on the minority class would give less attention to the majority class. However, the cost-sensitive learning algorithms achieved a much bigger goal by improving the correct predictions in the minority class. Lastly, except for the accuracy, an increase in the other performance evaluation metrics is observed in all the cost-sensitive algorithms compared to the cost-insensitive algorithms. Meanwhile, several research works on imbalance classification for medical diagnosis agree that it is more dangerous to misclassify a positive patient than misclassify a negative patient [10,67].

6. Conclusion

This research studied cost-sensitive learning and developed some cost-sensitive algorithms by modifying their loss function to focus more on the minority class. Three repeats of 10-fold cross-validation was used during the training of the various algorithms. The research utilized performance metrics such as AUC, accuracy, precision, recall, F-measure, and Cohen's Kappa coefficient to evaluate the performance of the classifiers. The experimental results showed that the cost-sensitive versions of random forest, XGBoost, and logistic regression obtained excellent performance in the four datasets compared to other algorithms and some recently proposed research works. The results obtained in this research demonstrates the potential of cost-sensitive learning in predicting imbalanced medical data. Future research works would focus on further improving the prediction of the minority class while also ensuring the algorithm does not neglect the majority class in the process. Future research works could also employ a hybrid approach by combining cost-sensitive learning and resampling techniques such as SMOTE and adaptive synthetic sampling techniques and comparing the performance with instances where cost-sensitive learning and resampling techniques are used individually. Another potential future research direction is the combination of feature selection, resampling, and cost-sensitive learning methods.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Acknowledgments

This work was supported in part by the South African National Research Foundation under Grant 120106 and Grant 132797, and in part by the South African National Research Foundation Incentive under Grant 132159.

References

[1] Casalino G, Castellano G, Zaza G. A mHealth solution for contact-less self-monitoring of blood oxygen saturation. In: In 2020 IEEE symposium on computers and communications. ISCC); Jul. 2020. p. 1–7. https://doi.org/10.1109/ISCC50000.2020.9219718.
[2] Ghorbani R, Ghousi R, Makui A, Atashi A. A new hybrid predictive model to predict the early mortality risk in intensive care units on a highly imbalanced dataset. IEEE Access 2020;8:141066–79. https://doi.org/10.1109/ACCESS.2020.3013320.
[3] Liu C, Hsieh P. Model-based synthetic sampling for imbalanced data. IEEE Trans Knowl Data Eng Aug. 2020;32(8):1543–56. https://doi.org/10.1109/TKDE.2019.2905559.
[4] Gajowniczek K, Ząbkowski T. ImbTreeEntropy and ImbTreeAUC: novel R packages for decision tree learning on the imbalanced datasets. Electronics Jan. 2021;10(6). https://doi.org/10.3390/electronics10060657. Art. no. 6.
[5] Liu N, Li X, Qi E, Xu M, Li L, Gao B. A novel ensemble learning paradigm for medical diagnosis with imbalanced data. IEEE Access 2020;8:171263–80. https://doi.org/10.1109/ACCESS.2020.3014362.
[6] Khan SH, Hayat M, Bennamoun M, Sohel FA, Togneri R. Cost-sensitive learning of deep feature representations from imbalanced data. IEEE Transactions on Neural Networks and Learning Systems Aug. 2018;29(8):3573–87. https://doi.org/10.1109/TNNLS.2017.2732482.
[7] Yang Q, Wu X. 10 challenging problems in data mining research. Int J Inf Technol Decis Making Dec. 2006;5(4):597–604. https://doi.org/10.1142/S0219622006002258.
[8] Thai-Nghe N, Gantner Z, Schmidt-Thieme L. Cost-sensitive learning methods for imbalanced data. In: The 2010 international joint Conference on neural networks. IJCNN); Jul. 2010. p. 1–8. https://doi.org/10.1109/IJCNN.2010.5596486.
[9] Zhang L, Zhang D. Evolutionary cost-sensitive extreme learning machine. IEEE Transactions on Neural Networks and Learning Systems Dec. 2017;28(12):3045–60. https://doi.org/10.1109/TNNLS.2016.2607757.
[10] He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng Sep. 2009;21(9):1263–84. https://doi.org/10.1109/TKDE.2008.239.
[11] Zhang C, Tan KC, Li H, Hong GS. A cost-sensitive deep belief network for imbalanced classification. IEEE Transactions on Neural Networks and Learning Systems Jan. 2019;30(1):109–22. https://doi.org/10.1109/TNNLS.2018.2832648.
[12] Jing X-Y, et al. Multiset feature learning for highly imbalanced data classification. IEEE Trans Pattern Anal Mach Intell Jan. 2021;43(1):139–56. https://doi.org/10.1109/TPAMI.2019.2929166.
[13] Masnadi-Shirazi H, Vasconcelos N. Cost-sensitive boosting. IEEE Trans Pattern Anal Mach Intell Feb. 2011;33(2):294–309. https://doi.org/10.1109/TPAMI.2010.71.
[14] Yu H, Sun C, Yang X, Zheng S, Wang Q, Xi X. LW-ELM: a fast and flexible cost-sensitive learning framework for classifying imbalanced data. IEEE Access 2018;6:28488–500. https://doi.org/10.1109/ACCESS.2018.2839340.
[15] Hoens TR, Chawla NV. Imbalanced datasets: from sampling to classifiers. In: Imbalanced learning. John Wiley & Sons, Ltd; 2013. p. 43–59. https://doi.org/10.1002/9781118646106.ch3.
[16] Ma Y, Zhao K, Wang Q, Tian Y. Incremental cost-sensitive support vector machine with linear-exponential loss. IEEE Access 2020;8:149899–914. https://doi.org/10.1109/ACCESS.2020.3015954.
[17] Balasubramanian S, Kashyap R, Cvn ST, Anuradha M. Hybrid prediction model for type-2 diabetes with class imbalance,. In: 2020 IEEE international conference on machine learning and applied network technologies (ICMLANT); Dec. 2020. p. 1–6. https://doi.org/10.1109/ICMLANT50963.2020.9355975.
[18] Xiaolong X, Wen C, Yanfei S. Over-sampling algorithm for imbalanced data classification. J Syst Eng Electron Dec. 2019;30(6):1182–91. https://doi.org/10.21629/JSEE.2019.06.12.
[19] Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. Journal of Big Data Mar. 2019;6(1):27. https://doi.org/10.1186/s40537-019-0192-5.
[20] Kuo RJ, Su PY, Zulvia FE, Lin CC. "Integrating cluster analysis with granular computing for imbalanced data classification problem – a case study on prostate cancer prognosis. Comput Ind Eng Nov. 2018;125:319–32. https://doi.org/10.1016/j.cie.2018.08.031.
[21] Liu T, Fan W, Wu C. A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. Artif Intell Med Nov. 2019;101:101723. https://doi.org/10.1016/j.artmed.2019.101723.
[22] Fernando KRM, Tsokos CP. Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks. IEEE Transactions on Neural Networks and Learning Systems 2021:1–12. https://doi.org/10.1109/TNNLS.2020.3047335.
[23] Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinf Mar. 2013;14(1):106. https://doi.org/10.1186/1471-2105-14-106.
[24] Zeng M, Zou B, Wei F, Liu X, Wang L. "Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. In: 2016 IEEE international conference of online analysis and computing science. ICOACS); May 2016. p. 225–8. https://doi.org/10.1109/ICOACS.2016.7563084.
[25] Xu Z, Shen D, Nie T, Kou Y. A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. J Biomed Inf Jul. 2020;107:103465. https://doi.org/10.1016/j.jbi.2020.103465.
[26] Shilaskar S, Ghatol A, Chatur P. Medical decision support system for extremely imbalanced datasets. Inf Sci Apr. 2017;384:205–19. https://doi.org/10.1016/j.ins.2016.08.077.
[27] Mienye ID, Sun Y, Wang Z. An improved ensemble learning approach for the prediction of heart disease risk. Informatics in Medicine Unlocked Jan. 2020;vol. 20:100402. https://doi.org/10.1016/j.imu.2020.100402.
[28] Zhang Y, Wang X, Han N, Zhao R. Ensemble learning based postpartum hemorrhage diagnosis for 5G remote healthcare. IEEE Access 2021;9:18538–48. https://doi.org/10.1109/ACCESS.2021.3051215.
[29] Zhu M, Su B, Ning G. Research of medical high-dimensional imbalanced data classification ensemble feature selection algorithm with random forest. " May 2017. p. 273–7. https://doi.org/10.1109/ICSGEA.2017.158.
[30] Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) Jul. 2012;42(4):463–84. https://doi.org/10.1109/TSMCC.2011.2161285.
[31] Zhou X, Li Y, Liang W. CNN-RNN based intelligent recommendation for online medical pre-diagnosis support. IEEE ACM Trans Comput Biol Bioinf 2020:1-1. https://doi.org/10.1109/TCBB.2020.2994780.
[32] Yaghoobi H, Babaei E, Hussen BM, Emami A. EBST: an evolutionary multi-objective optimization based tool for discovering potential biomarkers in ovarian cancer. IEEE ACM Trans Comput Biol Bioinf 2020:1-1. https://doi.org/10.1109/TCBB.2020.2993150.
[33] Mienye ID, Sun Y, Wang Z. Improved sparse autoencoder based artificial neural network approach for prediction of heart disease. Informatics in Medicine Unlocked Jan. 2020;18:100307. https://doi.org/10.1016/j.imu.2020.100307.
[34] Ali SI, et al. Ensemble feature ranking for cost-based non-overlapping groups: a case study of chronic kidney disease diagnosis in developing countries. IEEE Access 2020;8:215623–48. https://doi.org/10.1109/ACCESS.2020.3040650.
[35] Phankokkruad M. "Cost-Sensitive extreme gradient boosting for imbalanced classification of breast cancer diagnosis," in *2020 10th IEEE international Conference on control system*. Computing and Engineering (ICCSCE) Aug. 2020:46–51. https://doi.org/10.1109/ICCSCE50387.2020.9204948.
[36] Lomax S, Vadera S. A cost-sensitive decision tree learning algorithm based on a multi-armed bandit framework. Comput J Jul. 2017;60(7):941–56. https://doi.org/10.1093/comjnl/bxw015.
[37] Zięba M, Tomczak JM, Lubicz M, Świątek J. Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. Appl Soft Comput Jan. 2014;14:99–108. https://doi.org/10.1016/j.asoc.2013.07.016.
[38] Ali S, Majid A, Javed SG, Sattar M. Can-CSC-GBE: developing Cost-sensitive Classifier with Gentleboost Ensemble for breast cancer classification using protein amino acids and imbalanced data. Comput Biol Med Jun. 2016;73:38–46. https://doi.org/10.1016/j.compbiomed.2016.04.002.
[39] Wan X, Liu J, Cheung WK, Tong T. Learning to improve medical decision making from imbalanced data without a priori cost. BMC Med Inf Decis Making Dec. 2014;14(1):111. https://doi.org/10.1186/s12911-014-0111-9.
[40] Zhu M, et al. Class weights random forest algorithm for processing class imbalanced medical data. IEEE Access 2018;6:4641–52. https://doi.org/10.1109/ACCESS.2018.2789428.
[41] Gan D, Shen J, An B, Xu M, Liu N. Integrating TANBN with cost sensitive classification algorithm for imbalanced data in medical diagnosis. Comput Ind Eng Feb. 2020;140:106266. https://doi.org/10.1016/j.cie.2019.106266.
[42] Wang H, Cui Z, Chen Y, Avidan M, Abdallah AB, Kronzer A. Predicting hospital readmission via cost-sensitive deep learning. IEEE ACM Trans Comput Biol Bioinf Dec. 2018;15(6):1968–78. https://doi.org/10.1109/TCBB.2018.2827029.
[43] Wu J-C, Shen J, Xu M, Liu F-S. An evolutionary self-organizing cost-sensitive radial basis function neural network to deal with imbalanced data in medical diagnosis. Int J Comput Intell Syst Oct. 2020;13(1):1608–18. https://doi.org/10.2991/ijcis.d.201012.005.
[44] Ding J, Errapotu SM, Guo Y, Zhang H, Yuan D, Pan M. Private empirical risk minimization with analytic Gaussian mechanism for healthcare system. In: IEEE Transactions on big data; 2020. 1-1. https://doi.org/10.1109/TBDATA.2020.2997732.
[45] Makki S, Assaghir Z, Taher Y, Haque R, Hacid M, Zeineddine H. An experimental study with imbalanced classification approaches for credit card fraud detection. IEEE Access 2019;7:93010–22. https://doi.org/10.1109/ACCESS.2019.2927266.
[46] Jiang X, Pan S, Long G, Xiong F, Jiang J, Zhang C. Cost-sensitive parallel learning framework for insurance intelligence operation. IEEE Trans Ind Electron Dec. 2019;66(12):9713–23. https://doi.org/10.1109/TIE.2018.2873526.

[47] Ling CX, Sheng VS. Cost-sensitive learning. In: Sammut C, Webb GI, editors. Encyclopedia of machine learning. Boston, MA: Springer US; 2010. p. 231–5. https://doi.org/10.1007/978-0-387-30164-8_181.
[48] Lin Y. A note on margin-based loss functions in classification. Stat Probab Lett Jun. 2004;68(1):73–82. https://doi.org/10.1016/j.spl.2004.03.002.
[49] Yang Y, Huang S, Huang W, Chang X. Privacy-preserving cost-sensitive learning. " *IEEE Transactions on Neural Networks and Learning Systems*; 2020. p. 1–12. https://doi.org/10.1109/TNNLS.2020.2996972.
[50] Theodoridis S. In: Theodoridis S, editor. "Chapter 7 - classification: a tour of the classics," in *machine learning*. second ed. Academic Press; 2020. p. 301–50. https://doi.org/10.1016/B978-0-12-818803-3.00016-7.
[51] Fitriyani NL, Syafrudin M, Alfian G, Rhee J. HDPM: an effective heart disease prediction model for a clinical decision support system. IEEE Access 2020;8:133034–50. https://doi.org/10.1109/ACCESS.2020.3010511.
[52] Zhao X, Wu Y, Lee DL, Cui W. iForest: interpreting random forests via visual analytics. IEEE Trans Visual Comput Graph Jan. 2019;25(1):407–16. https://doi.org/10.1109/TVCG.2018.2864475.
[53] Pima Indians diabetes database. https://kaggle.com/uciml/pima-indians-diabetes-database. [Accessed 17 April 2021].
[54] UCI machine learning repository: Haberman's survival data set. https://archive.ics.uci.edu/ml/datasets/haberman%27s+survival. [Accessed 15 April 2021].
[55] Uci Machine Learning Repository: Cervical cancer (risk factors) data set. https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29. [Accessed 15 April 2021].
[56] UCI machine learning repository: chronic_Kidney_Disease data set. https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease. [Accessed 20 July 2021].
[57] Fernandes K, Cardoso JS, Fernandes J. "Transfer learning with partial observability applied to cervical cancer screening,". In: Pattern Recognition and image analysis, Cham; 2017. p. 243–50. https://doi.org/10.1007/978-3-319-58838-4_27.
[58] Kuhn M, Johnson K. Applied predictive modeling. first ed. New York: Springer; 2013. Corr. 2nd printing 2018 edition.
[59] Soleymani R, Granger E, Fumera G. F-measure curves: a tool to visualize classifier performance under imbalance. Pattern Recogn Apr. 2020;100:107146. https://doi.org/10.1016/j.patcog.2019.107146.
[60] Gárate-Escamila AK, Hajjam El Hassani A, Andrès E. Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked Jan. 2020;19:100330. https://doi.org/10.1016/j.imu.2020.100330.
[61] Ferri C, Hernández-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. Pattern Recogn Lett Jan. 2009;30(1):27–38. https://doi.org/10.1016/j.patrec.2008.08.010.
[62] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas Apr. 1960;20(1):37–46. https://doi.org/10.1177/001316446002000104.
[63] Chicco D, Tötsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min Feb. 2021;14(1):13. https://doi.org/10.1186/s13040-021-00244-z.
[64] Khan B, Naseem R, Muhammad F, Abbas G, Kim S. An empirical evaluation of machine learning techniques for chronic kidney disease prophecy. IEEE Access 2020;vol. 8:55012–22. https://doi.org/10.1109/ACCESS.2020.2981689.
[65] Wardhani NWS, Rochayani MY, Iriany A, Sulistyono AD, Lestantyo P. Cross-validation metrics for evaluating classification performance on imbalanced data,. In: *2019 international Conference on computer, control, Informatics and its applications (IC3INA)*; Oct. 2019. p. 14–8. https://doi.org/10.1109/IC3INA48034.2019.8949568.
[66] Mienye ID, Ainah PK, Emmanuel ID, Esenogho E. Sparse noise minimization in image classification using Genetic Algorithm and DenseNet,". In: 2021 Conference on information communications Technology and society. ICTAS); Mar. 2021. p. 103–8. https://doi.org/10.1109/ICTAS50802.2021.9395014.
[67] Branco P, Torgo L, Ribeiro R. A survey of predictive modelling under imbalanced distributions," *arXiv:1505.01658 [cs]*. May 2015. Accessed: Jan. 18, 2021. [Online]. Available: http://arxiv.org/abs/1505.01658.
[68] McHugh ML. Interrater reliability: the kappa statistic. Biochem Med Oct. 2012;22(3):276–82.
[69] Asniar NU, Maulidevi, Surendro K. SMOTE-LOF for noise identification in imbalanced data classification. Journal of King Saud University - Computer and Information Sciences Feb. 2021. https://doi.org/10.1016/j.jksuci.2021.01.014.
[70] Chatrati SP, et al. Smart home health monitoring system for predicting type 2 diabetes and hypertension. Journal of King Saud University - Computer and Information Sciences Jan. 2020. https://doi.org/10.1016/j.jksuci.2020.01.010.
[71] Hayashi Y, Yukita S. Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset. Informatics in Medicine Unlocked Jan. 2016;2:92–104. https://doi.org/10.1016/j.imu.2016.02.001.
[72] Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. " *ICT Express*; Feb. 2021. https://doi.org/10.1016/j.icte.2021.02.004.
[73] S. Wei, X. Zhao, and C. Miao, "A comprehensive exploration to the machine learning techniques for diabetes identification," in 2018 *IEEE 4th world Forum on Internet of things (WF-IoT)*, feb. 2018, pp. 291–295. doi: 10.1109/WF-IoT.2018.8355130.
[74] Syed AH, Khan T. Machine learning-based application for predicting risk of type 2 diabetes mellitus (T2DM) in Saudi arabia: a retrospective cross-sectional study. IEEE Access 2020;8:199539–61. https://doi.org/10.1109/ACCESS.2020.3035026.
[75] Abd El-Salam SM, et al. Performance of machine learning approaches on prediction of esophageal varices for Egyptian chronic hepatitis C patients. Informatics in Medicine Unlocked Jan. 2019;17:100267. https://doi.org/10.1016/j.imu.2019.100267.
[76] Pranto B, Mehnaz Sk M, Momen S, Huq SM. Prediction of diabetes using cost sensitive learning and oversampling techniques on Bangladeshi and Indian female patients,". In: 2020 5th international Conference on information technology research (ICITR); Dec. 2020. p. 1–6. https://doi.org/10.1109/ICITR51448.2020.9310892.
[77] Gorunescu F, Belciug S. Evolutionary strategy to develop learning-based decision systems. Application to breast cancer and liver fibrosis stadialization. J Biomed Inf Jun. 2014;49:112–8. https://doi.org/10.1016/j.jbi.2014.02.001.
[78] Kaushik D, Prasad BR, Sonbhadra SK, Agarwal S. "Post-Surgical survival forecasting of breast cancer patient: a novel approach," in *2018 international Conference on Advances in computing*. Communications and Informatics (ICACCI) Sep. 2018:37–41. https://doi.org/10.1109/ICACCI.2018.8554745.
[79] Aljawad DA, et al. Breast cancer surgery survivability prediction using bayesian network and support vector machines. In: 2017 international Conference on informatics, health technology (ICIHT); Feb. 2017. p. 1–6. https://doi.org/10.1109/ICIHT.2017.7899000.
[80] Cahyana N, Khomsah S, Aribowo AS. Improving imbalanced dataset classification using oversampling and gradient boosting,". In: *2019 5th international Conference on Science in information technology (ICSITech)*,; Oct. 2019. p. 217–22. https://doi.org/10.1109/ICSITech46713.2019.8987499.
[81] Ijaz MF, Attique M, Son Y. Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. Sensors Jan. 2020;20(10). https://doi.org/10.3390/s20102809. Art. no. 10.
[82] Ebiaredoh-Mienye SA, Esenogho E, Swart TG. Integrating enhanced sparse autoencoder-based artificial neural network technique and softmax regression for medical diagnosis. Electronics Nov. 2020;9(11). https://doi.org/10.3390/electronics9111963. Art. no. 11.
[83] Nithya B, Ilango V. Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction. SN Appl. Sci. May 2019;1(6):641. https://doi.org/10.1007/s42452-019-0645-7.
[84] Wu W, Zhou H. Data-driven diagnosis of cervical cancer with support vector machine-based approaches. IEEE Access 2017;5:25189–95. https://doi.org/10.1109/ACCESS.2017.2763984.
[85] Abdoh SF, Abo Rizka M, Maghraby FA. Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. IEEE Access 2018;6:59475–85. https://doi.org/10.1109/ACCESS.2018.2874063.
[86] Rashed-Al-Mahfuz M, Haque A, Azad A, Alyami SA, Quinn JMW, Moni MA. Clinically applicable machine learning approaches to identify attributes of Chronic Kidney Disease (CKD) for use in low-cost diagnostic screening. IEEE Journal of Translational Engineering in Health and Medicine 2021:1-1. https://doi.org/10.1109/JTEHM.2021.3073629.
[87] Ogunleye A, Wang Q-G. XGBoost model for chronic kidney disease diagnosis. IEEE ACM Trans Comput Biol Bioinf Nov. 2020;17(6):2131–40. https://doi.org/10.1109/TCBB.2019.2911071.
[88] Chittora P, et al. Prediction of chronic kidney disease - a machine learning perspective. IEEE Access 2021;9:17312–34. https://doi.org/10.1109/ACCESS.2021.3053763.
[89] Almustafa KM. Prediction of chronic kidney disease using different classification algorithms. Informatics in Medicine Unlocked Jan. 2021;24:100631. https://doi.org/10.1016/j.imu.2021.100631.