RESEARCH ARTICLE

DOI: 10.54392/irjmt2435

Asian Research Association

INTERNATIONAL RESEARCH JOURNAL OF
MULTIDISCIPLINARY TECHNOVATION

Enhanced Classification of Imbalanced Medical Datasets using Hybrid Data-Level, Cost-Sensitive and Ensemble Methods

Ayushi Gupta [a], Shikha Gupta [a, *]

[a] Department of Computer Science, Shaheed Sukhdev College of Business Studies, University of Delhi, Sector 16, Rohini, Delhi, India.

* Corresponding Author Email: shikhagupta@sscbsdu.ac.in

DOI: https://doi.org/10.54392/irjmt2435

Received: 17-01-2024; Revised: 18-02-2024; Accepted: 22-02-2024; Published: 22-04-2024

Abstract: Addressing the class imbalance in classification problems is particularly challenging, especially in the context of medical datasets where misclassifying minority class samples can have significant repercussions. This study is dedicated to mitigating class imbalance in medical datasets by employing a hybrid approach that combines data-level, cost-sensitive, and ensemble methods. Through an assessment of the performance, measured by AUC-ROC values, Sensitivity, F1-Score, and G-Mean of 20 data-level and four cost-sensitive models on seventeen medical datasets - 12 small and five large, a hybridized model, SMOTE-RF-CS-LR has been devised. This model integrates the Synthetic Minority Oversampling Technique (SMOTE), the ensemble classifier Random Forest (RF), and the Cost-Sensitive Logistic Regression (CS-LR). Upon testing the hybridized model on diverse imbalanced ratios, it demonstrated remarkable performance, achieving outstanding performance values on the majority of the datasets. Further examination of the model's training duration and time complexity revealed its efficiency, taking less than a second to train on each small dataset. Consequently, the proposed hybridized model not only proves to be time-efficient but also exhibits robust capabilities in handling class imbalance, yielding outstanding classification results in the context of medical datasets.

Keywords: Class Imbalance, Medical Data, Voting Ensemble, Machine Learning

1. Introduction

Class imbalance, an ongoing concern in machine learning, occurs when data points are unevenly distributed among different groups or classes. This imbalance poses challenges in classification problems and may introduce bias in the results [1]. Data samples that are fewer in number compared to others are referred to as minority class samples, whereas the remaining samples are termed majority class samples. During model training, limited data availability can hinder generalization to new, unseen samples from the minority class, leading to inaccuracies in predictions. If not appropriately penalized, models may overfit and optimize predominantly for the majority class samples [2]. In the process of feature engineering as well, the model might assign more importance to features relevant to the majority class.

Additionally, class imbalance renders models sensitive to changes in data distribution. Even slight alterations in the representation of the minority class can significantly impact model performance [3]. This is particularly crucial for achieving accurate predictions. Hence, effective handling of the class imbalance problem is essential to enhance the overall performance of machine learning models.

Furthermore, addressing the class imbalance is imperative to generate models that excel across all classes in the dataset. The issue is magnified in applications where the minority class corresponds to rare events, such as cancer detection in medical datasets. In such cases, misclassifying an instance can have severe consequences, including potential threats to life, for example - identifying a person having HIV falsely as negative.

1.1 Recent Literature

In recent times, numerous research studies have concentrated on addressing the challenge of class imbalance and have undertaken a comprehensive review of existing techniques in this domain. In the work by, the authors conducted experimental evaluations, analyzing the performance of boosting methods on both normal-sized (fifteen datasets) and large-scale (four datasets) datasets with imbalanced classes [4]. Their findings identified CatBoost and SMOTEBoost as the top performers based on the mean Matthews Correlation Coefficient over binary class pairs. Notably, the study revealed that oversampling methods surpass undersampling methods in the classification of imbalanced data. A more recent survey presented an evaluation strategy of 24 state-of-the-art techniques for imbalanced data streams [5]. Turning to the challenges and trends in addressing the class imbalance, highlighted the obstacles posed by this problem [6]. They explored publication trends, the popularity of different approaches, and the usage of tools in recent publications.

In terms of proposing innovative techniques, introduced a hybrid approach employing simulated annealing for undersampling and four diverse classifiers [7]. The method was evaluated on 51 real datasets, demonstrating superior performance compared to three other studies. The authors in the study, developed a hybrid parameterization model, utilizing soft set theory to reduce the number of parameters for neural network-based classification [8]. Feng *et al.* gave a hybrid algorithm for binary imbalanced datasets, incorporating Negative-positive Synthetic Minority Oversampling Technique (NPSMOTE), Binary Ant Lion Optimizer, and General Vector Machine, outperforming state-of-the-art methods across seven benchmark datasets [9]. A comparison of various resampling techniques using the random forest as the classifier and analyzed the performance of the High School Longitudinal Study of 2009 dataset has been presented in literature [10]. Their findings indicated that the hybrid technique SMOTE-NC and RUS combined, works best since random oversampling may lead to overfitting issues and random under sampling may lead to loss of useful information. In, the authors proposed a hybridized SMOTE algorithm combined with a genetic algorithm for optimized sampling [11]. They evaluated their technique on the Cup 1999 big dataset by dividing it into four instances with varying imbalance ratios and using a decision tree algorithm as the classifier. In, the authors proposed an undersampling technique based on K-Means and C-Means clustering approaches by replacing the sample clusters with the cluster heads [12]. They evaluated their approach on 12 big datasets using three different classifiers and the AUC evaluation measure.

1.2 Applications in Medical Domain

In this section, we discuss some applications of the Class imbalance methods in the medical domain. In the investigation the focus was on utilizing transfer learning and deep learning techniques for skin cancer detection using image data [13]. Addressing the challenge of a highly imbalanced dataset, they employed F1-score and AUC-ROC curves as performance evaluation metrics for the algorithms. In the study, three healthcare datasets were examined, and stacked deep learning models were analyzed in conjunction with two SMOTE-based resampling techniques to manage class imbalance [14]. The results demonstrated substantial accuracy improvements with the stacked architectures compared to contemporary machine learning algorithms. In the research, the Pima Indian Diabetes dataset was evaluated with a diabetes prediction model [15]. The methodology involved outlier removal, imputation of missing values, feature selection, and classification using k-nearest neighbors. Notably, the study did not specifically address the class imbalance issue despite working with an imbalanced dataset. In the research work the authors conducted a study on COVID-19 detection, analyzing chest X-ray images [16]. The images underwent conversion to vectors using an autoencoder, followed by resampling using standard techniques to achieve a balanced dataset. Subsequently, different state-of-the-art classifiers were employed for training and testing on the balanced data. In the study, the authors conducted a systematic survey on diabetic retinopathy detection and then extended their work on early detection of Diabetic Retinopathy using image data by modifying existing transfer learning models and utilizing upsampling for data balancing [17]. They also conducted a systematic survey on retinal imaging techniques for Alzheimer's disease detection [17]. In the study, the same authors proposed a deep learning ensemble model for retinal blood vessel segmentation in two public databases comprising fundus images [18].

1.3 Our Contributions

The present study concentrates on introducing hybrid techniques designed to address data imbalance challenges and improve classification tasks within medical datasets. In the present work, we have focused on: 1. Hybrid classifiers considering data-level, cost-sensitive, and ensemble methods. 2. Voting ensemble combining all three imbalance handling techniques. 3. A significant number of record-based medical datasets with varying imbalance ratios in the medical domain. 4. Multiple performance evaluation metrics to assess the models. The primary contributions of this study include:

I. Addressing class imbalance in medical datasets through the application of data-level, cost-sensitive, and ensemble methods.

II. Classifying 17 datasets in the medical domain including five large binary medical datasets using a set of 24 hybridized models.

III. Classification of binary datasets exhibiting diverse imbalanced ratios ranging from 1.13 to 54.45.

IV. Assessing the performance of the models using the relevant evaluation metrics - AUC-ROC, Sensitivity, F1-Score, and G-Mean.

V. Identification of the top two performers and proposing a voting ensemble hybridised model combining all three techniques for managing class imbalance and effective predictions on all the datasets.

VI. Reporting the training duration of the hybridised model on all the datasets and analysing its time complexity.

The rest of the paper is structured as follows: Section 2 outlines the popular class imbalance handling techniques. Section 3 presents the methodology employed in the present research. Section 4 presents the obtained results, followed by a detailed discussion. The conclusion and future directions are articulated in Section 5.

2. Class Imbalance Handling Techniques

The categorization of techniques for addressing class imbalance in datasets encompasses two primary types: data-level methods and algorithm-level methods. Data-level methods operate within the data space during pre-processing to achieve balance. Various data resampling techniques have been introduced in recent literature to address dataset imbalances, including over-sampling, under-sampling, and a combination of both. The data-level balancing techniques employed in this study are outlined below:

- Random Over Sampling (ROS) - This method achieves balance by randomly duplicating minority samples in the data space [19]. However, the generation of duplicate data may lead to overfitting issues, especially in small-sized datasets.
- Synthetic Minority Oversampling Technique (SMOTE) - Another widely used technique, SMOTE generates synthetic minority samples by employing a nearest neighbor strategy [20]. A new minority data sample is created on the virtual line segment between a minority feature vector and its nearest neighbor, iteratively repeated until the number of minority samples equals that of the majority class.
- Random Under Sampling (RUS) - This technique achieves balance by randomly removing majority class samples in the data space [21]. However, data deletion may result in the loss of important information and significantly reduce dataset size.
- Edited Nearest Neighbor (ENN) - This undersampling technique calculates the three nearest neighbors for each data sample [22]. If a majority class sample is misclassified according to its neighbors, it is removed from the dataset. Conversely, if a minority class sample is misclassified, all its majority neighbors are removed.
- SMOTE-ENN - This combined over and under sampling technique utilizes SMOTE to balance data samples and subsequently removes misclassified samples from both classes using ENN [23].
- Algorithm-level techniques operate on established classification algorithms, adapting them to exhibit a bias toward minority-class data samples and are classified as cost-sensitive and ensemble approaches. Cost-sensitive methods assign a greater cost to the misclassification of minority-class samples than to majority-class samples. In contrast, ensemble methods employ bagging and boosting techniques in conjunction with data-level or cost-sensitive methods to address the challenge of class imbalance. The algorithm-level methods employed in the present study are summarized as follows:
- ***Cost Sensitive Logistic Regression (CSLR)*** - It is a statistical method designed for binary classification tasks [24]. The term “logistic” is derived from its utilization of the logistic or sigmoid function to model the probability of a specific outcome. Let X_{tr} and y_{tr} denote the training set features and class labels, while X_{te} and y_{te} denote the test set features and class labels. The class predictions on the test set are denoted by y_{pr} . The misclassification cost associated with the test samples can be expressed as:

$$cost = -y_{te}\log(y_{pr}) - (1 - y_{te})\log(1 - y_{pr}) \qquad (1)$$

To introduce cost-sensitivity to minority samples, the algorithm’s $class\_weight$ parameter is set to *balanced*, which automatically adjusts weights based on the class labels, inversely proportional to their frequencies in the input data. To understand it better, let us have $m = 500$ samples in the data, having $nc = 2$ class labels with $|c_1| = 200$ samples belonging to the minority class and $|c_2| = 300$ samples belonging to the majority class. The class weights for c_1 and c_2 are then calculated as:

$$weights = \left[\frac{m}{nc*[200,300]}\right] = \left[\frac{500}{[400,600]}\right] \qquad (2)$$

$$weights = [1.25,0.83] \qquad (3)$$

These weights are then used to modify the misclassification cost associated with minority and majority samples, by heavily penalizing the minority sample misclassification.

- ***Cost Sensitive Decision Tree (CS-DT)*** - This method recursively divides the data into subsets, selecting the most significant feature at each step [25]. The structure consists of nodes, each representing a decision or test on a specific feature, interconnected by branches leading to other nodes or leaves. The uppermost node signifies the initial decision or test on the most crucial feature, and the terminal nodes provide the final decision or output. In scenarios with imbalanced class distributions, decision trees may exhibit bias toward the dominant class. To address this, cost sensitivity is incorporated into the classifier by configuring the $class\_weight$ parameter to *balanced*.

- ***Cost Sensitive Support Vector Machine (CS-SVM)*** - It is a powerful classification algorithm that is capable of constructing hyperplanes or decision boundaries that optimize the segregation between distinct classes [26]. SVM strategically selects the optimal hyperplane by maximizing the distance between training examples, with the closest points to the decision boundary termed support vectors. The emphasis on achieving a large margin stem from the notion that a more considerable distance between the hyperplane and support vector points results in better classification for points on either side of the plane. This classifier is therefore referred to as a **large margin classifier**. Once again, in addressing the class imbalance, the classifier incorporates cost-sensitivity by configuring the $class\_weight$ parameter to *balanced*.
- ***Cost Sensitive Extreme Gradient Boosting (CS-XGB)*** - This method systematically introduces weak learners (Decision Trees or DTs) to the model in a sequential manner, addressing errors made by preceding learners [27]. Initially, every data sample is accorded equal importance. Subsequently, upon analyzing prediction outcomes, greater emphasis is placed on incorrectly classified samples, and the data is reintroduced to the subsequent learner in the sequence. This iterative process continues until the misclassification error falls below a predefined threshold [28]. To enhance efficiency and optimize speed, the algorithm leverages parallel computing and distributed computing. To manage class imbalance, the $scale\_pos\_weight$ parameter is adjusted to the imbalanced ratio. This parameter denotes the cost associated with falsely classifying minority samples. When set to 1, equal weight is assigned to both minority and majority classes.
- ***Balanced Random Forest (RF)*** - This technique is an ensemble learning method that combines the predictions of multiple individual DTs to enhance overall performance [29]. Each tree is trained on a randomly selected subset of the training data, with each node in the tree considering a randomized subset of features for splitting. This is done to diminish the variance and correlation between the trees, which is high particularly when they are constructed using the same dataset. The final result is calculated as the mode or majority voting of the classes for classification tasks or the average prediction for regression tasks. To handle class imbalance, RF is coupled with one of the data-level resampling methods, resulting in a hybrid classifier that combines data-level and ensemble strategies.

3. Our Proposal and Experiments

This section introduces the research diagram that delineates the flow of the current study (see Figure 1). Our proposal can be explained as follows:

1. Initially, seventeen publicly available medical datasets are acquired from the UCI and Kaggle repositories (Section 3.1).
2. The datasets undergo pre-processing, involving the conversion of categorical attributes to numerical counterparts using label encoding. Additionally, datasets with missing values are addressed by removing instances containing such values.
3. The pre-processed datasets are then fed into machine learning algorithms employing class imbalance handling techniques, utilizing data-level or cost-sensitive methods (Section 2). Five data-level handling methods are utilized, including two undersampling techniques (RUS and ENN), two oversampling techniques (ROS and SMOTE), and one combined technique- SMOTE-ENN. The cost-sensitive approach involves setting specific algorithm parameters to incorporate a higher cost for misclassifying minority data samples.
4. Five machine learning algorithms (LR, DT, SVM, RF, and XGB) are hybridized with class imbalance handling techniques, resulting in 24 hybridized classification models which are - ROS-LR, RUS-LR, SMOTE-LR, ENN-LR, SMOTE-ENN-LR, ROS-DT, RUS-DT, SMOTE-DT, ENN-DT, SMOTE-ENN-DT, ROS-RF, RUS-RF, SMOTE-RF, ENN-RF, SMOTE-ENN-RF, ROS-SVM, RUS-SVM, SMOTE-SVM, ENN-SVM, SMOTE-ENN-SVM, CS-LR, CS-DT, CS-SVM, and CS-XGB.
5. For dividing the datasets into training and test sets, five-fold cross-validation has been employed.

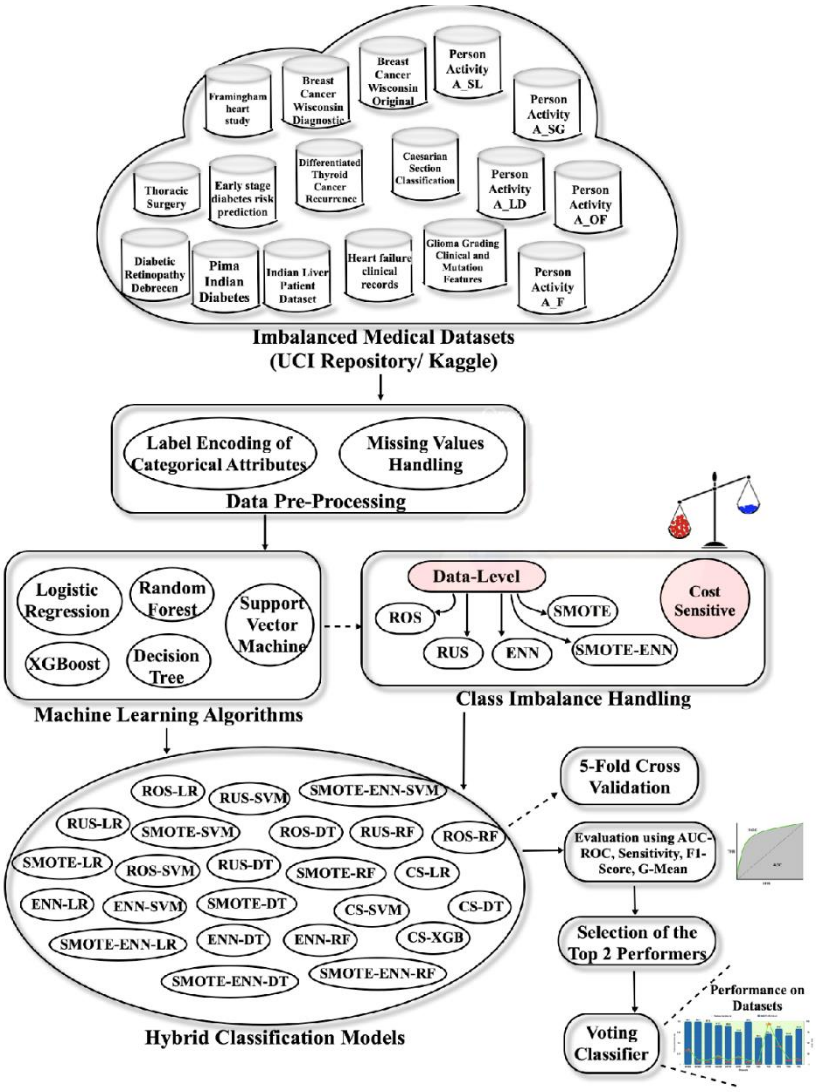


Figure 1. Research Diagram

6. These models are evaluated for performance on all the medical datasets using four evaluation metrics suitable for imbalanced datasets -AUC-ROC, Sensitivity, F1-Score, and G-Mean (Section 3.2).
7. After analyzing the performance of the models, the top two performing models are selected and further hybridized using a voting ensemble classifier (Section 3.3). The results of the voting classifier are then reported on all seventeen datasets.

3.1 Datasets

The current investigation employs seventeen publicly available medical datasets - 12 small and five large, sourced from Kaggle and the UCI repository, as outlined in Table 1. The table provides information on the total number of data samples and features for each dataset. In Table 1, IR denotes the Imbalance Ratio, calculated as the ratio of the number of majority samples to the number of minority samples in binary data.

It's worth noting that the BCWO dataset initially had 699 samples. However, due to the presence of missing data, pre-processing was performed, resulting in the removal of rows with missing values and a total of 683 samples in the processed dataset. Similarly, the original sample counts for the ILPD and FHS datasets were 583 and 4240, respectively. All samples containing missing values were removed, yielding 579 and 3658 samples, respectively. Additionally, it is important to highlight that, except for the BCWO, PID, and FHS datasets, the other datasets are either recently added to the public UCI repository and have thus been utilized in only a few studies or have not been employed in any previous studies.

Furthermore, the last five large datasets pertain to person activity and have been derived from the Localization Data for Person Activity dataset, which is initially a multiclass dataset. To suit the objectives of the present study, these datasets have been transformed into five binary datasets, allowing for the exploration of various imbalance ratios. Additionally, within the original dataset, two attributes—timestamp and date—are identical. Given that the data is collected for a single date, only the timestamp attribute is retained in the pre-processed data. Moreover, the values within the timestamp attribute have been normalized to a range between 0 and 1, and the categorical attributes — sequence and tag identifier—have been encoded in the preprocessed dataset for analytical purposes.

Table 1. Data description and Properties

S.No	Dataset	Description	Rows	Features	IR*
1	Breast Cancer Wisconsin (Diagnostic) - BCWD [30]	Based on cell nucleus attributes like radius, texture, perimeter, concavity, etc. the data samples are grouped into Malignant and Benign	569	30	1.68
2	Breast Cancer Wisconsin (Original) - BCWO [31]	Based on attributes like clump thickness, cell properties, and nuclei properties, the data samples are grouped into Malignant and Benign	683	9	1.86
3	Differentiated Thyroid Cancer Recurrence – DTCR [32]	Collected over 15 years, the dataset comprises of clinicopathologic attributes for predicting the recurrence of thyroid cancer	383	16	2.55
4	Glioma Grading Clinical and Mutation Features – GGCM [33]	Based on 20 genes and 3 clinical features, the patients are categorized to be affected with Lower-Grade Glioma or Glioblastoma Multiforme- the two forms of the primary tumor in brain	839	23	1.38
5	Heart failure clinical records – HFCR [34]	Based on clinical features like age, diabetes, platelets, sex, etc. the patients are grouped based on whether they had a heart failure or not	299	12	2.11
6	Indian Liver Patient Dataset – ILPD [35]	Based on attributes like age, gender, bilirubin, and albumin properties, etc. the patients are grouped based on whether they suffer from liver disease or not	579	10	2.5
7	Early stage diabetes risk prediction dataset – DRP [36]	Based on symptoms like sudden weight loss, weakness, visual blurring, itching, etc. the patients are predicted whether they are diabetic or not	520	17	1.6
8	Caesarian Section Classification Dataset – CSC [37]	Based on age, delivery type, blood pressure, etc. the delivery is predicted to be caesarian or not	80	5	1.35
9	Framingham heart study dataset – FHS [38]	Based on smoking habits, blood pressure, diabetes, cholesterol, etc. the patients are grouped whether they possess a ten-year risk of future heart disease or not	3658	15	5.57
10	Diabetic Retinopathy Debrecen – DRD [39]	Based on the lesion and anatomical features, the patients are classified with signs of diabetic retinopathy or not	1151	19	1.13
11	Thoracic Surgery Data – TSD [40]	Based on before surgery symptoms, smoking habits, etc. whether the patients survived or not after one year of lung cancer operation	470	16	5.71
12	Pima Indian Diabetes - PID [41]	Based on diagnostic values such as number of pregnancies, BMI, insulin level, etc. whether the female patients have diabetes or not	768	8	1.87

13	Localization Data for Person Activity (Standing from Lying) - A_SL [42]	Based on tag data for ankles, belt, and chest, whether the person was standing up from lying down or not	164860	6	7.98
14	Localization Data for Person Activity (Sitting on Ground) - A_SG [42]	Based on tag data for ankles, belt, and chest, whether the person was sitting on the ground or not	164860	6	12.99
15	Localization Data for Person Activity (Lying Down) - A_LD [42]	Based on tag data for ankles, belt, and chest, whether the person was lying down or not	164860	6	25.73
16	Localization Data for Person Activity (On All Fours) - A_OF [42]	Based on tag data for ankles, belt, and chest, whether the person was on all fours or not	164860	6	30.64
17	Localization Data for Person Activity (Falling) - A_F [42]	Based on tag data for ankles, belt, and chest, whether the person was falling or not	164860	6	54.45

*IR: Imbalanced Ratio for binary data = $\frac{\#\ Majority\ Samples}{\#\ Minority\ Samples}$

The implementation was conducted utilizing the Python programming language with the Scikit-learn library and the Imblearn library [43-44]. The executions took place in an environment powered by macOS Big Sur Version 11.3.1, featuring 8GB of RAM.

3.2. Evaluation Metrics

The following evaluation metrics have been employed:

1. ***AUC ROC:*** Area Under the Curve – Receiver Operating Characteristics, AUC-ROC curve is a graphical representation that illustrates the trade-off between the true positive rate and the false positive rate and AUC represents the area under the ROC curve. An AUC value of 1 represents ideal classification while a 0.5 value indicates that the model is not performing better than a randomized version. Hence, this metric is suitable for assessing model performances on imbalanced datasets.

2. ***Sensitivity:*** Sensitivity, also referred to as the true positive rate (TPR) or recall, serves as a pivotal metric in assessing the efficacy of a machine learning model's ability to detect positive instances. This parameter quantifies the proportion of actual positives accurately identified by the model.

$$Sensitivity = \frac{TP}{TP+FN} \qquad (4)$$

3. ***F1-Score:*** The F1 score amalgamates precision and recall scores to offer a nuanced assessment of a model's accuracy. Computed through the harmonic mean of precision and recall, maximizing the F1 score necessitates the simultaneous optimization of both metrics, presenting a balanced evaluation of the model's effectiveness. Precision *P* discerns the true positive rate among records classified as positive, while Recall *R* identifies the fraction of true positive records among the actual positive records.

$$P = \frac{TP}{TP+FP} \qquad (5)$$

$$F1 - Score = 2 * \frac{P*R}{P+R} \qquad (6)$$

4. ***G Mean:*** Computed as the geometric mean of class-specific sensitivities, the Geometric Mean(G-mean) strives to enhance overall accuracy while preserving class balance. In binary classification, it is analogous to the square root of the product of sensitivity (true positive rate) and specificity (true negative rate). Traditionally, the G-mean yields zero when any class goes unrecognized.

$$Specificity = \frac{TN}{TN+FP} \qquad (7)$$

$$GMean = \sqrt{Sensitivity * Specificity} \qquad (8)$$

3.3 Voting Classifier

A Voting Classifier constitutes a machine learning paradigm characterized by its reliance on an ensemble of diverse models for training. The model's predictive output, denoting a specific class, is determined by the highest probability assigned to that class by the constituent models. This classifier aggregates the outcomes of each model incorporated into it, predicting the output class based on the predominant majority of votes.

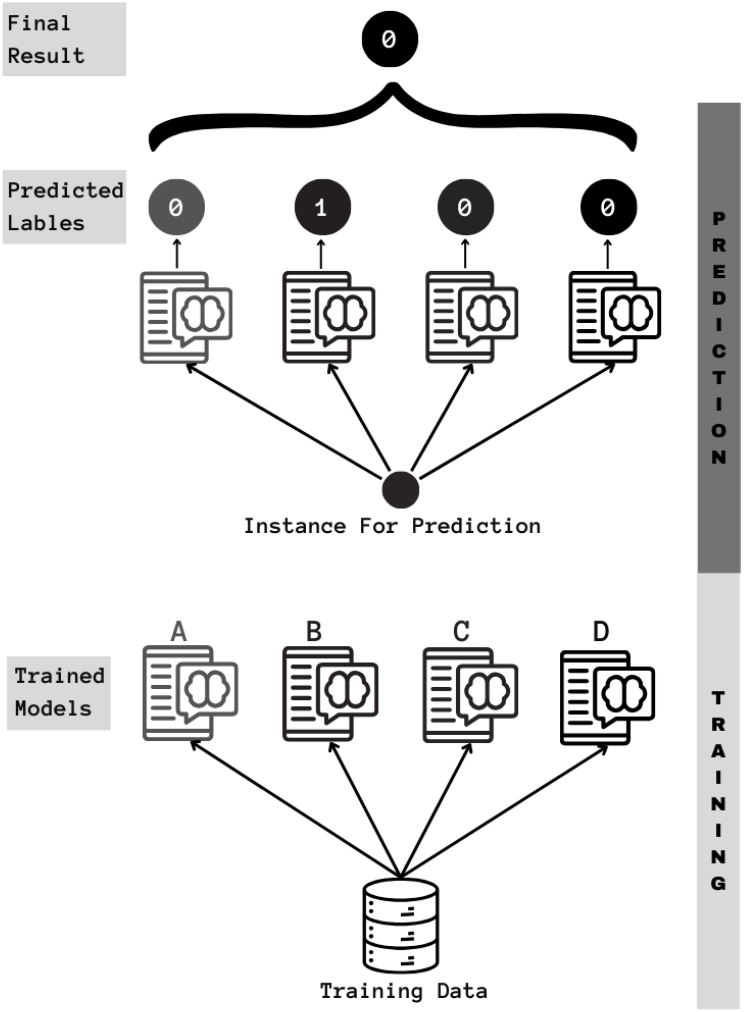


Figure 2. Voting Classifier Diagram

The underlying concept involves formulating a single model, acquiring training from multiple models, and predicting output through a consensus derived from their collective majority voting for each output class. The illustration in Figure 2 elucidates the functionality of voting classifiers.The depicted scenario involves the training of four distinct models denoted as A, B, C and D using the provided training data. Subsequently, these trained models collectively contribute to the prediction of the class for a given instance. In this specific instance, three out of the four classifiers collectively predict the label to be ′0′ leading to the final assignment of the label ′0′ for that particular instance.

4. Results and Discussion

The present section showcases the outcomes achieved through the application of the suggested hybrid models on the seventeen medical datasets. Table 1 displays the AUC-ROC values obtained by all the models across all datasets. Given the proximity of these values and their susceptibility to slight variations in different iterations, the table highlights the top three AUC-ROC values for each dataset.

Similarly, Table 3, Table 4 and Table 5 present the Sensitivity, F1-score, and G-mean values obtained on all the datasets with the top three values highlighted. Furthermore, these tables identify the top two performers in terms of both data-level and cost-sensitive models, based on their frequency in achieving the best performance (among the highest three values) across each metric. Key observations include:

- Considering AUC, none of the models utilizing Decision Trees (DT) for classification achieved the highest values for any of the datasets. However, SMOTE-DT was able to achieve good Sensitivity, F1-score, and G-mean solely on the CSC dataset and SMOTE-ENN-DT achieved good values on the TSD and the A_F large dataset.
- A similar trend is observed with Support Vector Machine (SVM) models, with only two models (ROS-SVM and SMOTE-SVM) exhibiting noteworthy performance, solely on the FHS dataset. The model SMOTE-ENN-SVM exhibits satisfactory performance only on the BCWO dataset.

Table 2. AUC ROC values obtained on the 17 medical datasets

CLASSIFIER	BCWD	BCWO	DTCR	GGCM	HFCR	ILPD	DRP	CSC	FHS	DRD	TSD	PID	A_SL	A_SG	A_LD	A_OF	A_F
ROS-LR	**99.2**	**99.5**	94.3	**91.5**	85.9	73.9	97.3	**65.6**	**72.4**	**82.8**	**65.7**	**83.3**	57.4	82.7	78.7	85.7	78.5
RUS-LR	99.1	**99.5**	94.5	**91.5**	86.4	74.8	97.2	60.8	71.8	**82.9**	59.1	83.1	57.4	82.6	78.6	85.7	78.5
SMOTE-LR	**99.2**	**99.5**	94.6	**91.5**	83.2	74.6	97.3	**66.6**	65.9	**82.7**	64.8	**83.4**	57.4	82.7	78.6	85.7	78.6
ENN-LR	99	**99.5**	95	90.2	85.3	**75.1**	96.6	44.8	**72.4**	81.1	**66.3**	82.6	57.2	80	78.4	85.3	78.5
SMOTE-ENN-LR	98.3	**99.5**	95	89.8	77.8	74.3	96.1	47.4	67	79.4	63.7	82.6	57.3	82.7	78.7	85.8	78.6
ROS-DT	92.6	96.2	93.2	89.3	77.7	69	96.4	58.6	66	66.6	51.4	78.7	72.1	87.5	83.5	89.6	83.9
RUS-DT	89.7	97	93.1	89.9	71.3	66.5	95.5	50	65.1	66.8	54.6	74.8	72.1	87.5	83	89.6	83.2
SMOTE-DT	90.2	96.7	93.3	89	75.8	67.3	97.2	58.8	66.6	67.2	57.3	76.2	72	87.5	83.7	89.9	85
ENN-DT	93.3	96	92.7	88.2	80.5	66.3	94.1	50	66.4	66	55.7	73.3	71.6	90.7	83.6	88.8	84.8
SMOTE-ENN-DT	93.5	95.3	92.9	86.9	69.3	67.7	92.6	56.4	65.7	66.4	58.4	76.9	72	87.5	83.7	89.8	84.8
ROS-SVM	97.1	98.8	92	82.7	45.4	73	73.6	44.5	**72.8**	77.9	58	81.1	59.8	**91.2**	79.4	89.8	82.5
RUS-SVM	97	98.9	84.1	82.2	47.5	72.3	67.4	57.6	71.5	77.6	53.8	80.5	58.2	84.1	78.8	87.1	78.1
SMOTE-SVM	97.2	98.6	91.4	82.6	48.8	72.7	73.3	44.9	**72.4**	77.8	51.6	81.2	59.9	91.1	79.6	89.9	82.7
ENN-SVM	97	98.5	89.2	81.9	44	72.5	72	58.2	70.3	78.3	55.9	79.9	55.5	90.6	58.3	79.6	70.6
SMOTE-ENN-SVM	97	98.7	92.1	81.7	45.2	73	83.1	45.8	72.1	73.8	59.9	79.9	59.3	91.1	79.5	89.9	82.6
ROS-RF	99.1	**99.2**	**98.1**	**90.5**	**90**	**75.4**	**99.9**	61.6	68.9	75.7	64.8	81.8	**90.4**	**99.7**	92.4	**97.5**	91
RUS-RF	**99.2**	99.1	**98.2**	**90.8**	89.7	72.9	**99.7**	63.6	69.7	75.6	59	81.6	88.5	**99.7**	90.9	96.8	90
SMOTE-RF	**99.3**	**99.2**	**98.1**	**90.8**	**90.4**	74.3	**99.9**	62.2	67.2	74.8	**66.3**	81.2	**90.9**	**99.8**	**93.3**	**97.9**	**92.7**
ENN-RF	99	98.9	97.9	89.6	88.7	**75.3**	99.2	59.5	70.5	74.2	65	81.6	**90.4**	**99.7**	**92.6**	**97.5**	**91.4**
SMOTE-ENN-RF	98.7	98.9	97.2	89.4	83.2	**75.1**	98.6	64	68.6	73.1	62.6	82.4	**90.7**	**99.7**	**93.4**	**97.9**	**92.8**
CS-LR	**99.2**	**99.5**	94.7	**91.5**	86.1	74.9	97.3	**65**	72	**82.9**	**65.6**	**83.2**	57.3	82.7	78.6	85.7	78.5
CS-DT	93	97.1	94.1	89.3	71.3	70.3	96.4	55.1	67.9	68.2	53.9	78.1	72.2	87.5	83.5	89.6	83.9
CS-SVM	97.1	98.6	89.2	82.5	48.8	72.9	69.7	49.6	**72.7**	77.7	53.8	81.1	59.4	87.8	79.2	89.2	81.5
CS-XGB	**99.5**	**99.3**	**98.7**	**90.5**	**90.1**	71.7	**99.6**	57.3	64.4	76.9	60.9	79.7	84	**99.7**	89.5	**96.9**	89

Table 3. Sensitivity values obtained on the 17 medical datasets

CLASSIFIER	BCWD	BCWO	DTCR	GGCM	HFCR	ILPD	DRP	CSC	FHS	DRD	TSD	PID	A_SL	A_SG	A_LD	A_OF	A_F
ROS-LR	95.1	96.8	88.2	**87.6**	76.6	68.7	92.7	**64.6**	**66.7**	74.5	59.8	**75.6**	55.5	80.2	75.7	78.1	74.9
RUS-LR	95.1	96.9	88.1	87.5	75.9	68.8	91.7	58.6	**66.6**	**75.2**	58.9	74.5	55.4	80.1	75.6	78.1	74.8
SMOTE-LR	95.2	96.9	87.5	**88**	73.7	**70.7**	92.1	**64.2**	62	**75.1**	**61.2**	**75.3**	55.5	80.2	75.6	78.2	74.6
ENN-LR	94.7	**97.5**	88.2	86.6	74.5	**71.4**	90.3	49	58.6	69.9	50.8	73.9	50	49.9	50	53.1	50
SMOTE-ENN-LR	92.2	97.3	88	86.5	69.6	70.2	90.1	46.3	61.2	72.6	59.5	72.8	56.1	80.2	76.2	78.4	75
ROS-DT	92.9	95.9	91.8	86.6	76.5	65.9	94.3	59.1	62.8	63.1	50.1	73.6	65.8	84.8	77	82.6	78.7
RUS-DT	90.6	96.2	93	87.2	71.4	67.2	92.3	53.3	63.9	62.8	52.8	72.4	65.7	84.8	76.8	83	78
SMOTE-DT	91.2	95.3	91.2	86.5	77.8	64.4	94.6	**64**	62.9	64.5	54	72.5	65.7	84.9	76.8	82.7	79.2
ENN-DT	92.8	95.2	92.3	86.5	79.9	66.1	92.8	51	56.1	62.6	53.1	72	57.5	59.6	50.2	59.1	50
SMOTE-ENN-DT	93.5	95.3	92	86.7	69.3	68.2	91.5	56.4	61.5	66	**60.8**	74.3	65.6	84.9	77.2	82.6	**79.3**
ROS-SVM	89.7	97	79.8	75.3	51.5	66.2	66.1	52.3	**66.5**	70.4	50.1	72.5	58	82.9	76.8	83.5	77.3
RUS-SVM	89.4	97	66.7	74.3	50.8	66.1	59.8	56.9	64.6	70	52.5	71.4	57.1	79.7	76.5	79.7	75.5
SMOTE-SVM	90	97.1	79.4	75.4	52.4	66.5	64.8	50.9	**66.7**	70.2	50.2	72.5	58.1	82.9	76.8	83.6	77.5
ENN-SVM	89.7	97.3	67.5	74.5	49.5	66.9	57.4	50	51.7	64.3	50	72.8	50	50	50	50	50
SMOTE-ENN-SVM	89.6	**97.6**	81.6	74.1	51.4	65.6	53.2	50.3	63.6	67.3	50.9	72.8	57.9	82.9	76.8	83.5	77.4
ROS-RF	**96.1**	**97.6**	**94.1**	83	**80**	61.2	**98.5**	58.3	54.8	69.2	53.3	72.1	70.7	96.4	60.3	76.6	55.4
RUS-RF	**96.1**	97.1	**93.4**	83.9	**81.1**	66.4	97.8	61.2	64.1	69.2	59.5	73.6	**80.4**	**97.6**	**83.3**	**91.6**	**82.7**
SMOTE-RF	**96.2**	97.3	**94.1**	84.2	**82.1**	64.9	**98.4**	55.1	55.6	69.1	52.5	72.9	**80**	**97.4**	79.8	88.9	76.1
ENN-RF	95.8	97.2	92.8	85.8	79	**72.1**	94.6	54.4	59.1	65.1	50.6	73.5	73.1	96.6	62.1	79	55.9
SMOTE-ENN-RF	93.8	**97.8**	91.9	86.7	72	68.5	93.8	56.4	61.9	67.3	58.2	**76.4**	**81.8**	**97.7**	**83**	**91.2**	**79.4**
CS-LR	95.2	96.7	87	**87.7**	76.9	69.9	92.4	59.9	66.4	**74.9**	**62.7**	75	55.4	80.2	75.7	78.1	74.9
CS-DT	93.7	96.2	93.2	86.2	74.7	68	95.4	59	63.4	65.5	53.3	72.5	65.7	84.8	77.2	82.5	78.6
CS-SVM	89.9	97.3	78.2	75	53	66.3	64.5	54.4	66	70.3	49.4	71.7	57.9	81.3	76.8	82.6	76.9
CS-XGB	**97**	96.3	**94.9**	81.7	79.5	62.2	**97.9**	52.5	56.3	69.6	53.6	71.4	75.9	**97.7**	**80.7**	**91**	77.8

Table 4. F1-Score values obtained on the 17 medical datasets

CLASSIFIER	BCWD	BCWO	DTCR	GGCM	HFCR	ILPD	DRP	CSC	FHS	DRD	TSD	PID	A_SL	A_SG	A_LD	A_OF	A_F
ROS-LR	96.4	95.6	80.9	**85.5**	67.7	55.8	93.6	**63.2**	**38.1**	72.9	30.8	**68.3**	21.7	32.4	16.4	18.4	8.3
RUS-LR	96.2	95.7	81.1	85.4	66.8	56	92.5	58.3	**37.8**	**73.5**	30.2	67.1	21.7	32.3	16.4	18.5	8.3
SMOTE-LR	96.5	95.8	80.2	**85.9**	64.3	**57.8**	93.3	61	33.5	**73.3**	**32.4**	68	21.7	32.5	16.5	18.5	8.3
ENN-LR	95.6	96	81.5	84.4	65.1	**58.5**	90.4	22.3	29.6	59	10.5	66.8	0	0	0.1	11	0.1
SMOTE-ENN-LR	94.8	95.7	81.4	84.3	59.1	57.2	90.3	31.2	32	68.2	30.1	65.5	22.2	32.3	16.3	18.2	8.1
ROS-DT	94.7	94.2	88.4	84.4	67.8	52.6	94.8	57.6	34.1	60.3	19.2	66	31.4	36.9	16.6	21.4	9.2
RUS-DT	92.8	94.3	88.8	85	61.3	54.1	93.7	46.1	35	58.2	24.5	64.6	33	37	16.2	19.7	9.1
SMOTE-DT	93.1	93.6	87.5	84.3	69.4	49.8	95.6	**64.2**	34.7	59.9	19.1	64.6	31.2	36.9	16.4	22.3	10.2
ENN-DT	93.6	93.4	87.2	84.3	71.5	52.8	92.9	27	24.4	47	18.6	65	25.3	29.8	0.9	26.4	0
SMOTE-ENN-DT	94.9	93.7	87.9	84.5	59.6	55.2	91.9	43.5	32.2	60.7	**31.5**	67	30.2	36.9	16.7	20.5	10.1
ROS-SVM	92.7	95.4	70.8	72.2	25.6	53.6	74.1	36.3	**37.8**	65.1	21.1	64.6	23.4	33.1	15.4	23.7	8.1
RUS-SVM	92.9	95.5	51.3	71.2	28.3	53.7	66.4	35.3	36.7	64.3	25.8	63.3	22.8	30.1	15.6	19.5	7.4
SMOTE-SVM	92.8	95.6	70	72.6	23.3	53.9	70.4	35.6	**37.7**	64.7	25.6	64.5	23.5	33.2	15.5	23.7	8.2
ENN-SVM	93	95.9	51.8	72.2	48.2	54.2	74.4	0	7.1	46.8	0	65.7	0	0	0	0	0
SMOTE-ENN-SVM	92.9	**96.2**	73	71.7	38.1	53.3	17.4	20.6	33.4	61	26.2	65.7	23.2	33.1	15.4	23.1	8.1
ROS-RF	**97**	**96.3**	**91.9**	80.2	**72.8**	44.2	**98.9**	**61.6**	19.4	68.9	14.7	63.4	52.6	**93.6**	30.4	**61**	18
RUS-RF	96.7	95.6	89.3	81.4	**73.6**	52.8	98.1	60.5	35.4	67.7	30.2	66	47.4	84.8	23.7	36.1	12.4
SMOTE-RF	**97.2**	95.9	**92**	81.6	**75.2**	49.8	**98.7**	60.8	23.1	68.7	12.1	64.7	**57.2**	**92.8**	**40.5**	**60.5**	**26.5**
ENN-RF	96.3	95.8	88.5	83.6	69.9	**59.1**	94.6	26.1	30.7	51	12.3	66.5	**54.3**	**93.1**	**33.5**	**61.4**	**19.6**
SMOTE-ENN-RF	95	**96.4**	87.3	84.5	62	55.2	94.5	38.9	33.7	63.7	29.9	**69.4**	**55**	91.5	**38**	57.2	**24.2**
CS-LR	96.4	95.4	79.7	**85.6**	68	56.9	93.4	58.4	**37.7**	**73.2**	**34**	**67.6**	21.7	32.4	16.5	18.4	8.3
CS-DT	95.6	94.5	89.8	83.9	65.6	54.9	95.8	59.3	35.2	61.4	25.3	64.8	31.4	36.9	16.5	21.2	9.3
CS-SVM	93	95.9	68.7	72.1	27.3	53.8	71.3	45	37.5	64.8	24.1	63.5	23.3	31.3	15.5	22.5	7.9
CS-XGB	**98.1**	94.8	**92.5**	78.8	72	45.4	**98.3**	58.1	25.4	71.1	19.2	62.6	43.2	88.5	25.1	45.8	15.1

Table 5. G-Mean values obtained on the 17 medical datasets

CLASSIFIER	BCWD	BCWO	DTCR	GGCM	HFCR	ILPD	DRP	CSC	FHS	DRD	TSD	PID	A_SL	A_SG	A_LD	A_OF	A_F
ROS-LR	95.1	96.8	88.2	**87.5**	76.5	66.9	92.7	**62.6**	**66.6**	74	58.3	**75.5**	55.5	79.9	75.3	78.1	74.4
RUS-LR	95	96.9	88	87.4	75.6	67.3	91.6	57.5	**66.5**	**74.6**	58.5	74.4	55.3	79.8	75.2	78.1	74.4
SMOTE-LR	95.2	96.9	87.4	**87.9**	73.6	**70**	92.1	**62.7**	61.8	**74.5**	**60.2**	**75.2**	55.4	79.9	75.2	78.2	74.2
ENN-LR	94.7	**97.4**	88.2	86.3	73.6	**69.7**	90.1	30.3	47.3	64.5	24.1	73	0	0	1.4	25.5	1.6
SMOTE-ENN-LR	92	97.2	88	86.3	68.4	68.7	89.9	33.9	59.2	70.9	56.9	72.4	54.6	79.8	75.5	78.4	74.4
ROS-DT	92.9	95.9	91.7	86.5	76	65.1	94.3	56.8	62.8	61.7	43.2	73.5	65.4	84.3	76.3	82.5	77.9
RUS-DT	90.6	96.1	92.9	87	71	66.3	92.3	49.9	63.6	61.2	49	72.1	64.1	84.3	75.8	82.7	77.3
SMOTE-DT	91.2	95.3	91	86.3	77.1	63.1	94.5	**63.3**	62.4	62.7	39.5	72.2	65.2	84.4	75.9	82.7	**78.8**
ENN-DT	92.7	95.2	92.2	86.3	**79.6**	65.7	92.6	33.8	42.5	54.9	37.6	71.1	42.6	44.9	6.5	43.4	0
SMOTE-ENN-DT	93.5	95.3	91.9	86.6	69.1	68.1	91.4	46	59.7	63.9	**60.3**	74.1	65.6	84.4	76.4	82.3	**78.9**
ROS-SVM	89.5	97	79	75.2	39.8	63.5	65.1	40.6	**66.4**	68.4	40.3	72.4	57.9	82	75	83.5	75.6
RUS-SVM	89.1	97	63.3	74.1	41.8	62.9	58.3	40.9	64.1	67.8	34.3	71.3	56.8	78.8	75.2	79.7	73
SMOTE-SVM	89.8	97.1	78.6	75.1	37.4	63.7	64.2	39.7	**66.6**	68.1	26.4	72.4	57.9	82.1	75.1	83.5	75.9
ENN-SVM	89.4	97.3	60.8	73.9	0	64.2	49.8	0	19.2	55.2	0	71.8	0	0	0	0	0
SMOTE-ENN-SVM	89.3	**97.6**	81.2	73.6	11.3	61.1	27.3	8.2	59	65	17.5	72	54.9	82	75	83.5	75.7
ROS-RF	**96.1**	**97.6**	**94**	82.9	79.4	58.4	**98.5**	57.7	34.7	69	26.9	71.3	65.7	96.4	45.9	73.3	33.1
RUS-RF	**96.1**	97.1	**93.4**	83.9	**81**	65.9	97.8	60.5	64	68.6	58.6	73.5	**80.4**	**97.6**	**83.1**	**91.6**	**82.5**
SMOTE-RF	**96.2**	97.3	**94**	84.1	**81.8**	63.2	**98.4**	53.4	40.7	68.8	22.7	72.5	**79.3**	**97.4**	78.5	88.6	73.7
ENN-RF	95.7	97.2	92.7	85.4	78.3	**71.1**	94.4	34.3	48.9	58.2	24.6	72.4	69.6	96.5	49.9	76.5	34.5
SMOTE-ENN-RF	93.8	**97.8**	91.9	86.5	70.6	68	93.8	42.8	61.3	66	55.9	**76.1**	**81.6**	**97.7**	**82.6**	**91.1**	78.2
CS-LR	95.2	96.7	87	**87.6**	76.7	68.3	92.3	58.5	**66.4**	**74.3**	**61.8**	74.9	55.4	79.8	75.3	78.1	74.4
CS-DT	93.6	96.2	93.1	86	74.3	67	95.3	58.3	63.2	64.1	49.2	72.4	65.1	84.3	76.2	82.4	77.8
CS-SVM	89.6	97.3	77.2	74.9	41.7	63.3	63.9	33.5	65.9	68.2	34.2	71.5	57.6	80.3	75.2	82.6	75.1
CS-XGB	**96.9**	96.3	**94.8**	81.6	79.1	59.4	**97.9**	50.5	45	69.4	37.9	70.6	75.7	**97.7**	**80.7**	**91**	77.3

- The oversampling data-level LR models (ROS-LR and SMOTE-LR) as well as its cost-sensitive variant (CS-LR), achieved top values across quite a few datasets. Notably, **CS-LR** demonstrated exceptional results on seven datasets in terms of AUC.

- Undersampling ensemble models such as RUS-RF and ENN-RF attained favorable outcomes for most large datasets with higher imbalance ratios. However, this achievement stems from the significant reduction in data size by these undersampling models, rendering the scale of the dataset irrelevant.

- Ensemble models based on oversampling, such as ROS-RF and SMOTE-RF, achieved outstanding results on the majority of the datasets (both small and large). Notably, SMOTE-RF emerged as the top model for achieving good performance (among the best three values) on 12 datasets based on AUC, six datasets for both Sensitivity and G-mean, and nine datasets for F1-score. Similarly, the model ROS-RF is ranked as the second top model for attaining good outcomes on nine, five, eight, and four datasets concerning AUC, Sensitivity, F1-score, and G-mean, respectively.

- Across all the tables, it is evident that undersampling and combined data-level techniques underperform in the majority of the cases. Notably, the SMOTE oversampling techniques hybridized with classification algorithms, particularly SMOTE-RF, yield promising results across most datasets. Additionally, cost-sensitive techniques, such as CS-LR and CS-XGB, demonstrate strong performance across the majority of datasets, securing the top two positions in the cost-sensitive paradigm. Also, CS-LR can perform well on FHS, DRD, TSD, and PID datasets where most of the other models are underperforming.

4.1 Discussion

Considering the top-performing models, specifically SMOTE-RF (data-level), CS-LR, and CS-XGB (cost-sensitive), two hybrid models employing a voting ensemble technique have been devised to achieve favorable outcomes across all datasets. The initial model, denoted as **SMOTE-RF-CS-LR**, combines the SMOTE data-level approach with cost-sensitive LR classification and an RF ensemble. Conversely, the second model, labeled **SMOTE-RF-CS-XGB**, integrates the SMOTE data-level method with cost-sensitive XGB classification and an RF ensemble. The performance of these models across all seventeen medical datasets is depicted in Figure 3. The following observations can be made:

1 Both the proposed hybrid models consistently demonstrate comparable or superior performance values compared to the best values in Table 2, Table 3, Table 4 and Table 5.

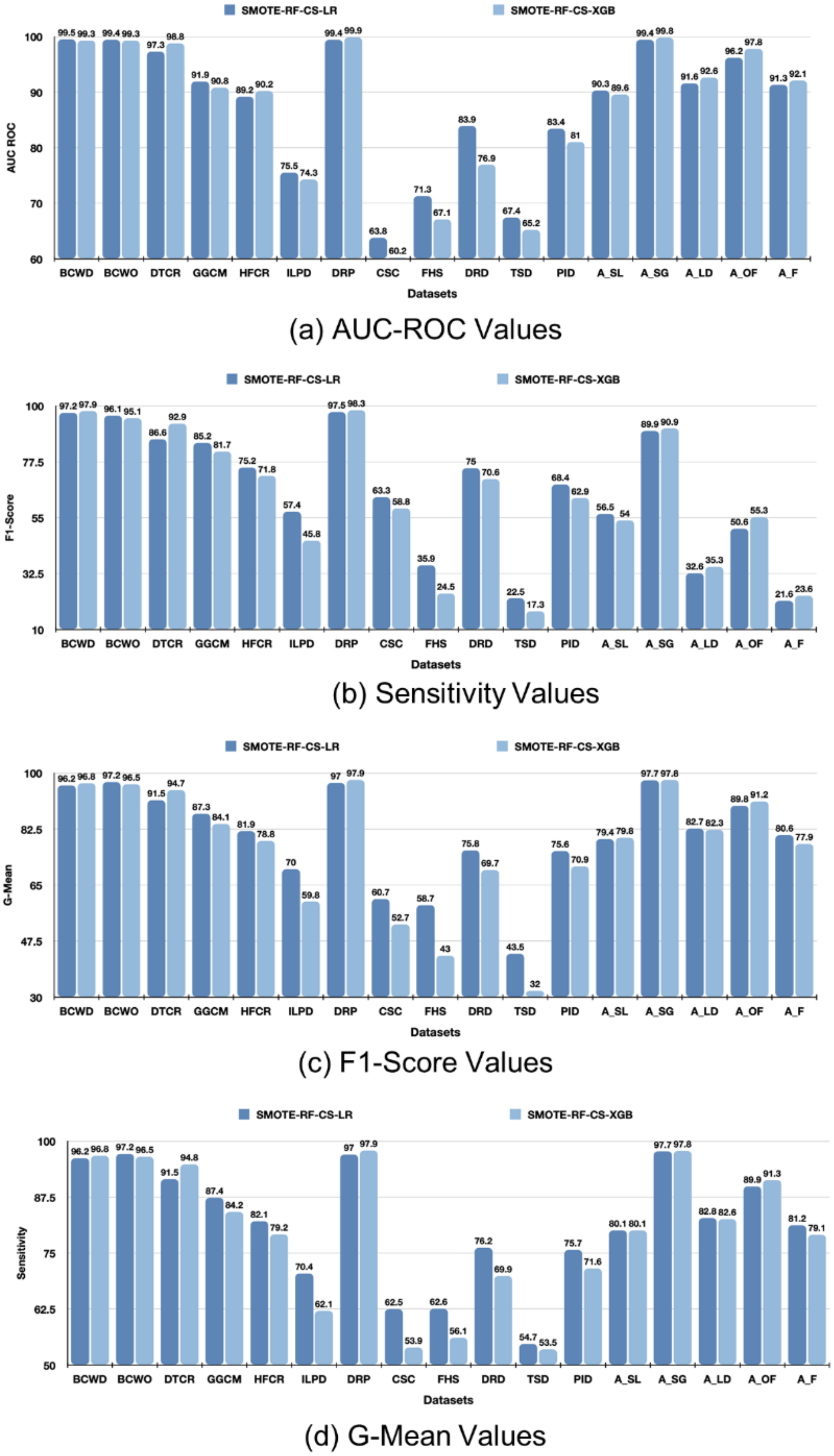
2 Among the two models, SMOTE-RF-CS-LR outperforms SMOTE-RF-CS-XGB on eight small datasets - BCWO, GGCM, ILPD, CSC, FHS, DRD, TSD, and PID, across all performance metrics - AUC ROC, Sensitivity, F1-Score, and G-Mean. Additionally, on the HFCR dataset, SMOTE-RF-CS-LR exhibits superior performance in three out of fours metrics. Furthermore, on three large datasets - A_SL, A_LD and A_F, it demonstrates comparable or better performance as compared to SMOTE-RF-CS-XGB.

3 Conversely, SMOTE-RF-CS-XGB surpasses SMOTE-RF-CS-LR in all metrics only on two small datasets - DTCR and DRP. It marginally improves over SMOTE-RF-CS-LR for the two large datasets - A_SG and A_OF and for the smaller BCWD dataset.

In Figure 4, the line graphs depict the models' training duration (in seconds) across all datasets. Upon analysis, it is evident that the model SMOTE-RF-CS-LR consistently required less time for training on each dataset when compared to SMOTE-RF-CS-XGB, indicating its notable speed.

Time Complexity Analysis: Delving into the time complexity of the proposed model SMOTE-RF-CS-LR, it is derived by summing up the individual time complexities of SMOTE, RF, and CS-LR. For SMOTE, the time complexity is given by $O(n\ log_2 n)$ [47], for RF, it is given by $O(n_t * n_v * n \log n)$ where n_t is the number of trees, n_v is the number of features and n is the number of data points. Since n_t and n_v are very small compared to n, they can be treated as constants and the complexity is reduced to $O(n \log n)$. For CS-LR, it is given by $O(n_v * n \log n)$, where again n_v can be treated as a constant. Therefore, the overall time complexity of the model SMOTE-RF-CS-LR is given by $O(n \log n)$.

All the above observations state that the hybridized model SMOTE-RF-CS-LR turns out to be the optimal choice for binary classification in imbalanced medical datasets. The favorable outcomes obtained by the model SMOTE-RF-CS-LR can be attributed to the efficacy of the SMOTE oversampling technique, which surpasses ROS by synthesizing minority samples rather than merely duplicating them. This approach helps mitigate overfitting issues. Additionally, the RF ensemble technique, known for its robust classification capabilities and reduced variance in DTs, achieves this by leveraging only essential features and demonstrating resistance from outliers.



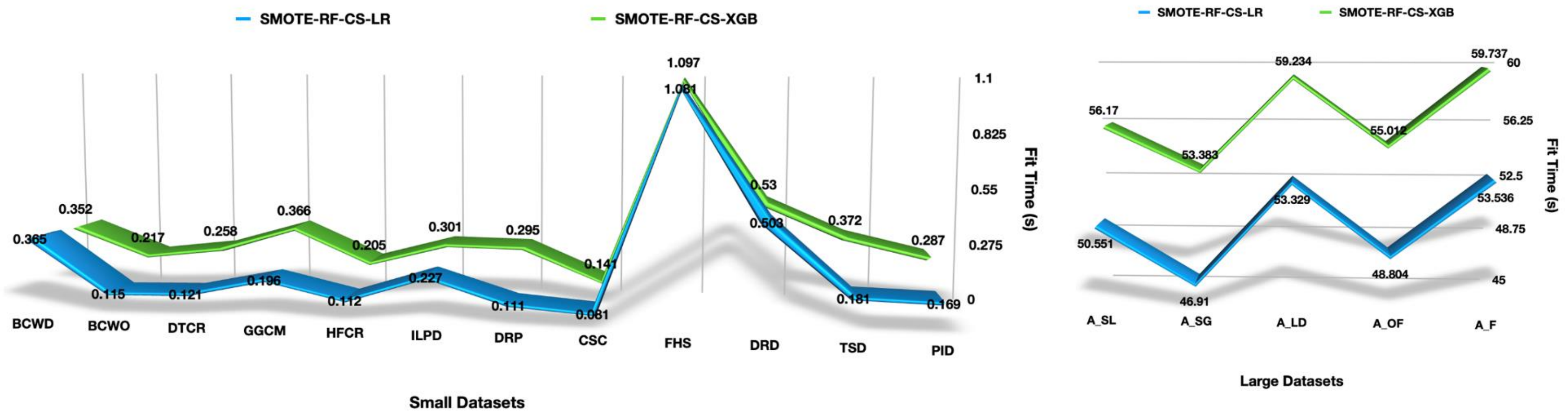
(a) AUC-ROC Values

(b) Sensitivity Values

(c) F1-Score Values

(d) G-Mean Values

Figure 3. Perfomance of Hybrid Voting Classifiers on all the Datasets



(a) Fit Time taken by Voting Classifiers on Small Datasets

(b) Fit Time taken by Voting Classifiers on Large Datasets

Figure 4. Fit Time taken by Hybrid Voting Classifiers on all the Datasets

Table 6. Comparison of SMOTE-RF-CS-LR with the Baseline Models on the 17 Medical Datasets

CLASSIFIER	BCWD	BCWO	DTCR	GGCM	HFCR	ILPD	DRP	CSC	FHS	DRD	TSD	PID	A_SL	A_SG	A_LD	A_OF	A_F
AUC ROC																	
DT	92.1	96.7	95.2	89	76.7	65.4	96.8	56	67.5	67.3	49.8	77.7	71.5	90.6	83.5	89	84.2
LR	99.2	99.5	94.6	91.6	86.2	74.7	97.5	65.9	71.3	82.9	65.6	83.3	57.3	80	78.4	85.4	78.5
SVM	97.3	99	91	82.8	52.5	69	90.1	70.6	62.4	77.2	51.6	81.1	54.2	90.5	58.7	67.7	71.4
RF	99.2	99.3	98.3	90.7	89.2	73.1	99.9	60.6	70.6	75	69.4	82.1	90.1	99.7	92.2	97.4	90.2
SMOTE-RF-CS-LR	**99.5**	**99.4**	97.3	**91.9**	89.2	**75.5**	99.4	63.8	71.3	**83.9**	**67.4**	**83.4**	90.3	**99.4**	91.6	96.2	91.3
Sensitivity																	
DT	93.3	94.6	94	86.9	72.8	54.2	94.5	62.9	52.8	64.5	51	69.7	50.7	59.5	50.1	52.3	50

LR	94.8	96.2	87	87.3	77	55.7	92.6	63.5	52.4	74.8	50.5	73.1	50	49.9	50	51.3	50
SVM	90	97.2	62.4	73.3	50	50	50	50	50.2	70.6	50	70	50	50	50	50	50
RF	95.9	97.3	93.2	82.9	78.9	57	98.4	56.2	52	68.3	49.1	69.5	65.3	95.3	55.6	70.5	52.6
SMOTE-RF-CS-LR	**96.2**	97.2	91.5	87.4	**82.1**	70.4	97	62.5	62.6	**76.2**	54.7	**75.7**	**80.1**	**97.7**	**82.8**	89.9	**81.2**
F1-Score																	
DT	95.1	93	91.2	84.8	63.1	28.9	95.2	62.9	13	61.3	11.2	59.7	3.2	29.7	0.4	7.5	0
LR	96.4	95.2	81.8	85.2	68.9	28.6	94.2	69	11	73.8	4.4	64.4	0	0	0	5	0
SVM	94	95.9	40	68.2	0	0	76.2	73	0.7	66.9	0	59.2	0	0	0	0	0
RF	97.2	96.1	91	80.1	71.7	33.9	98.7	62.7	9.3	68.5	0	59.3	44.6	93	19.6	53.7	9.9
SMOTE-RF-CS-LR	**97.2**	96.1	86.6	85.2	**75.2**	57.4	97.5	**63.3**	35.9	**75**	22.5	**68.4**	**56.5**	89.9	32.6	50.6	**21.6**
G-Mean																	
DT	93.2	94.5	93.9	86.8	71.5	43.6	94.4	62.2	27.4	63.1	24.4	68.1	12.6	44.6	4.4	13.6	0
LR	94.7	96.1	86.8	87.3	75.5	43.2	92.6	58.6	24.8	74.4	10.5	71.5	0	0	0.6	16.3	0
SVM	89.4	97.2	50.6	72.8	0	0	0	0	3.8	69.3	0	66.7	0	0	0	0	0
RF	95.9	97.3	93.1	82.8	78.1	48.5	98.4	55.1	22.6	68.1	0	67.8	56.3	95.2	33.8	64.1	23
SMOTE-RF-CS-LR	**96.2**	97.2	91.5	87.3	**81.9**	**70**	97	60.7	58.7	**75.8**	43.5	**75.6**	**79.4**	**97.7**	**82.7**	89.8	**80.6**

Further, by incorporating cost-sensitive LR, the model combines the robustness of LR in handling binary data with a specific emphasis on minimizing misclassifications of minority samples. As a result, the hybridized model proves to be robust and consistently delivers outstanding performances, particularly on imbalanced medical binary datasets.

4.2. Comparison with the Baseline Models

This subsection presents the comparison of the optimal hybridised model SMOTE-RF-CS-LR with four baseline models--- DT, LR SVM, and RF, in terms of all four performance metrics on all seventeen datasets. Table 6 depicts the comparison results. It is evident from the table that the proposed hybridised model SMOTE-RF-CS-LR is able to achieve best results across majority of the datasets in terms of all the four performance measures. Please note that the lower F1-Score values are due to the low precision values in the case of datasets possessing higher imbalance ratios. Also, the 0 values of F1-Score and G-Mean indicate that the model is unable to recognize one of the classes completely.

5. Conclusion

This study addresses the classification challenges within imbalanced binary medical datasets, employing both data-level and algorithm-level techniques. Five data-level methods, namely ROS, RUS, SMOTE, ENN, and SMOTE-ENN, are combined with four classifiers (LR, SVM, DT, and RF), resulting in the creation of 20 distinct models. Furthermore, the study explores four cost-sensitive models - CS-LR, CS-DT, CS-SVM, and CS-XGB.

The evaluation is conducted across 12 small and five large medical datasets, utilizing the AUC-ROC characteristic, Sensitivity, F1-Score, and G-Mean measures.

The findings reveal that the hybridization of the data-level method SMOTE with the RF ensemble (SMOTE-RF) and the incorporation of the cost-sensitive method into LR (CS-LR) and XGB (CS-XGB), consistently yield favorable performance across a majority of datasets. Motivated by these observations, two new hybridized models, SMOTE-RF-CS-LR and SMOTE-RF-CS-XGB, are introduced, integrating data-level SMOTE, ensemble classifier RF, and cost-sensitive LR and XGB methods. These novel models are tested across all datasets, demonstrating improved or comparable results to the best-performing models across all datasets. However, upon comparison, SMOTE-RF-CS-LR emerges as the optimal choice due to its superior performance across the majority of datasets (both small and large). Furthermore, analysis of the training durations reveals that SMOTE-RF-CS-LR consistently requires less time than SMOTE-RF-CS-XGB across all datasets.

In conclusion, the proposed hybridized model, incorporating various class imbalance handling methods, proves to be cost-efficient and capable of making accurate predictions on medical datasets. Nevertheless, the study has its limitations. It exclusively addresses the class imbalance problem within binary medical datasets. Future research could extend its focus to addressing class imbalance issues in multi-label medical datasets and explore applications across various domains. Also, several techniques for handling and imputing missing values can be investigated for improved performance. Furthermore, with better hardware, future research may concentrate on imbalance issues in big datasets, incorporating much higher imbalance ratios.

References

[1] N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study. Intelligent data analysis, 6(5), (2002) 429-449.

[2] A. Ali, S.M. Shamsuddin, A.L. Ralescu, Classification with class imbalance problem. International Journal of Advances in Soft Computing and its Applications, 5(3), (2013) 176–204.

[3] M.C. Monard, G. Batista, Learning with skewed class distributions. Advances in Logic, Artificial Intelligence and Robotics, 85, (2002) 173–180.

[4] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, M. Asadpour, Boosting methods for multi-class imbalanced data classification: an experimental review. Journal of Big Data, 7, (2020) 1–47. https://doi.org/10.1186/s40537-020-00349-y

[5] G. Aguiar, B. Krawczyk, A. Cano, A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework. Machine Learning, (2023) 1–79. https://doi.org/10.1007/s10994-023-06353-6

[6] P. Kaur, A. Gosain, Issues and challenges of class imbalance problem in classification. International Journal of Information Technology, 14(1), (2022) 539–545. https://doi.org/10.1007/s41870-018-0251-8

[7] A.S. Desuky, S. Hussain, An improved hybrid approach for handling class imbalance problem. Arabian Journal for Science and Engineering, 46, (2021) 3853–3864. https://doi.org/10.1007/s13369-021-05347-7

[8] M. Mohamad, A. Selamat, I.M. Subroto, O. Krejcar, Improving the classification performance on imbalanced data sets via new hybrid parameterisation model. Journal of King Saud University-Computer and Information Sciences, 33(7), (2021) 787–797. https://doi.org/10.1016/j.jksuci.2019.04.009

[9] F. Feng, K.C. Li, E. Yang, Q. Zhou, L. Han, A. Hussain, M. Cai, A novel oversampling and feature selection hybrid algorithm for imbalanced data classification. Multimedia Tools and Applications, 82(3), (2023) 3231–3267. https://doi.org/10.1007/s11042-022-13240-0

[10] D. Elreedy, A.F. Atiya, F. Kamalov, A theoretical distribution analysis of synthetic minority oversampling technique (smote) for imbalanced learning. Machine Learning, (2023) 1–21. https://doi.org/10.1007/s10994-022-06296-4

[11] M. Prince, P.J. Prathap, An imbalanced dataset and class overlapping classification model for big data. Computer Systems Science and Engineering, 44(2), (2023) 1009–1024. https://doi.org/10.32604/csse.2023.024277

[12] K. Ahlawat, A. Chug, A.P. Singh, Benchmarking framework for class imbalance problem using novel sampling approach for big data. International Journal of System Assurance Engineering and Management, 10, (2019) 824–835. https://doi.org/10.1007/s13198-019-00817-6

[13] A.S. Qureshi, T. Roos, Transfer learning with ensembles of deep neural networks for skin cancer detection in imbalanced data sets. Neural Processing Letters, 55(4), (2023) 4461–4479. https://doi.org/10.1007/s11063-022-11049-4

[14] A.M. Sowjanya, O. Mrudula, Effective treatment of imbalanced datasets in health care using modified smote coupled with stacked deep learning algorithms. Applied Nanoscience, 13(3), (2023) 1829–1840. https://doi.org/10.1007/s13204-021-02063-4

[15] M. kumari, P. Ahlawat, Dcpm: An effective and robust approach for diabetes classification and prediction. International Journal of Information Technology, 13, (2021) 1079–1088. https://doi.org/10.1007/s41870-021-00656-4

[16] S. Chatterjee, S. Maity, M. Bhattacharjee, S. Banerjee, A.K. Das, W. Ding, Variational autoencoder based imbalanced covid19 detection using chest x-ray images. New Generation Computing, 41(1), (2023) 25–60. https://doi.org/10.1007/s00354-022-00194-y

[17] R. Vij, S. Arora, A systematic review on diabetic retinopathy detection using deep learning techniques. Archives of Computational Methods in Engineering, 30(3), (2023) 2211–2256. https://doi.org/10.1007/s11831-022-09862-0

[18] R. Vij, & S. Arora, A hybrid evolutionary weighted ensemble of deep transfer learning models for retinal vessel segmentation and diabetic retinopathy detection. Computers and Electrical Engineering, 115, (2024) 109107. https://doi.org/10.1016/j.compeleceng.2024.109107

[19] C.X. Ling, C. Li, Data mining for direct marketing: Problems and solutions. In: Kdd, 98, (1998) 73–79

[20] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, (2002) 321–357. https://doi.org/10.1613/jair.953

[21] J. Prusa, T.M. Khoshgoftaar, D.J. Dittman, A. Napolitano, (2015) Using random undersampling to alleviate class imbalance on tweet sentiment data. IEEE International Conference on Information Reuse and Integration, IEEE. San Francisco. https://doi.org/10.1109/IRI.2015.39

[22] D.L. Wilson, Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics (3), (1972) 408–421. https://doi.org/10.1109/TSMC.1972.4309137

[23] G.E. Batista, R.C. Prati, M.C. Monard, A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 6(1), (2004) 20–29. https://doi.org/10.1145/1007730.1007735

[24] R.E. Wright, Logistic regression. Reading and understanding multivariate statistics, (1995) 217–244.

[25] J.R. Quinlan, Simplifying decision trees. International journal of man-machine studies, 27(3), (1987) 221–234. https://doi.org/10.1016/S0020-7373(87)80053-6

[26] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines. IEEE Intelligent Systems and their applications, 13(4), (1998) 18–28. https://doi.org/10.1109/5254.708428

[27] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system. Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, (2016) 785–794. https://doi.org/10.1145/2939672.2939785

[28] A. Gupta, A. Chug, and A. P. Singh, “Processing and optimized learning for improved classification of categorical plant disease datasets,” Intelligent Data Analysis, no. Preprint, pp. 1–25.

[29] L. Breiman, Random forests. Machine learning, 45(1), (2001) 5–32. https://doi.org/10.1023/A:1010933404324

[30] W. Wolberg, O. Mangasarian, N. Street, W. Street, (1995) Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository.

[31] W. Wolberg, (1992) Breast Cancer Wisconsin (Original). UCI Machine Learning Repository.

[32] Borzooei, S., Tarokhian, A.: Differentiated Thyroid Cancer Recurrence. UCI Machine Learning Repository. (2023)

[33] E. Tasci, K. Camphausen, A.V. Krauze, Y. Zhuge, (2022) Glioma Grading Clinical and Mutation Features. UCI Machine Learning Repository.

[34] T. Ahmad, A. Munir, S.H. Bhatti, M. Aftab, M.A. Raz, (2020) Heart Failure Clinical Records Dataset. UCI Machine Learning Repository.

[35] B. Ramana, N. Venkateswarlu, (2012) ILPD (Indian Liver Patient Dataset). UCI Machine Learning Repository. https://doi.org/10.24432/C5D02C

[36] Early stage diabetes risk prediction dataset. (2020) UCI Machine Learning Repository.

[37] M. Amin, A. Ali, (2018). Caesarian section classification dataset. UCI Machine Learning Repository.

[38] A. Bhardwaj, (2022) Framingham heart study dataset. Kaggle.

[39] B. Antal, A. Hajdu, (2014) Diabetic Retinopathy Debrecen. UCI Machine Learning Repository.

[40] M. Lubicz, K. Pawelczyk, A. Rzechonek, J. Kolodziej, (2013) Thoracic Surgery Data. UCI Machine Learning Repository.

[41] D. Dua, C. Graff, (2017) UCI Machine Learning Repository.

[42] V. Vidulin, M. Lustrek, B. Kaluza, R. Piltaver, J. Krivec. (2010) Localization Data for Person Activity. UCI Machine Learning Repository.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, (2011) 2825–2830.

[44] G. Lemaˆıtre, F. Nogueira, C.K. Aridas, Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 18(17), (2017) 1–5.

[45] X. Xiaolong, C. Wen, S. Yanfei, Oversampling algorithm for imbalanced data classification. Journal of Systems Engineering and Electronics 30(6), (2019) 1182–1191. https://doi.org/10.21629/JSEE.2019.06.12

Acknowledgements

We extend our heartfelt appreciation to Saurabh Ranjan, Gagan Rana, Sachin, and Vansh Shah for their exemplary efforts in conducting the literature survey for this study. We also thank Animesh Kumar for his help in compiling the data.

Authors Contribution Statement

Both the authors contributed to the study design and data collection. The structure of the work and the article was conceptualized by Dr. Shikha Gupta. The coding part and first draft of the manuscript were written by Ayushi Gupta and both authors commented on previous versions of the manuscript. Both authors read and approved the final manuscript. Dr. Shikha Gupta served as the corresponding author.

Funding

The study did not receive any funding.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

The datasets utilized in this study are publicly available at the UCI machine learning repository.

Has this article screened for similarity?

Yes

About the License