Received 2 August 2023, accepted 21 August 2023, date of publication 29 August 2023, date of current version 1 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3309732

RESEARCH ARTICLE

A New Multi-Class Rebalancing Framework for Imbalance Medical Data

JAFHATE EDWARD [1], MARSHIMA MOHD ROSLI [1,2], AND ALI SEMAN [1]

[1]College of Computing, Informatics and Mathematics, MARA University of Technology, Shah Alam, Selangor 40450, Malaysia

[2]Institute for Pathology, Laboratory and Forensic Medicine (I-PPerForM), MARA University of Technology (UiTM), Sungai Buloh, Shah Alam, Selangor 47000, Malaysia

Corresponding author: Marshima Mohd Rosli (marshima@uitm.edu.my)

ABSTRACT Class imbalance exists in many data domains, posing numerous challenges to the data research community. Medical datasets, in most cases, are predominantly imbalanced in nature. Through tackling multi-class issues, most researchers preferred the conventional method of decomposing it into binary classes for a more convenient solution. This method is not applicable for solving sensitive and crucial domains, such as medical data. Classifying medical datasets require all the classes to retain their form and maintain clinical validity. In this article, we develop a rebalancing framework for the multi-classification of imbalanced medical data using SCUT (SMOTE and Cluster-based Undersampling Technique) to rebalance the imbalanced class distribution, a feature selection method using a combination of SHapley Additive exPlanations (SHAP) and Recursive Feature Elimination (RFE), and DES-MI (Dynamic Ensemble Selection for multi-class) for improved multi classification performance. Two novelties contribute to the performance of our framework: improvised SCUT by implementing two clustering algorithms, and our proposed pool classifier selection for DES-MI. The performance of the proposed framework was compared with other state-of-the-art imbalanced frameworks using eight imbalanced datasets, each with varying degrees of imbalance. The experimental results indicate that our proposed framework performed better with average performance of 81.77%, 73.57%, and 75.87% in terms of Macro Average accuracy, extended G-mean, and Macro Average AUC, respectively. Our framework drastically increases the overall performance, owing to its ability to significantly handles the multi-class imbalance problem.

INDEX TERMS Imbalanced data, medical data, rebalancing framework, multi-class, classification prediction.

I. INTRODUCTION

Class imbalance emerges as one of the concerns that challenge many researchers in all data domains regardless of its application. The imbalance of classes can be defined as when one class (majority) outnumbers the instances of another class (minority) [1]. Exists in both binary and multi-class problems. Learning datasets that harbours this issue may hinder a model's predictive performance, resulting in a more biased model that favours the majority class and thus increases the misclassification rate.

Many researchers have engrossed their attention to overcoming this challenge, thus several rebalancing frameworks have been proposed in recent years [2], [3], [4], [5], [6], [7], [8], [9], [10]. Implementing various data and algorithm-level methods in a unified framework for learning medical imbalanced data, as such, their works yielded significant results. However, these works focus on the conventional decomposition method to solve multi-class problems. The decomposition method requires the transformation of multi-class into subproblems of binary class [11]. Binary classes are easier to solve since it only involves two classes (positive and negative). However, multi-classes are more complex because it includes subclasses of positive and negative classes [12]. Thus, most researchers favoured this decomposition method for its convenience [13], [14], [15], [16], [17]. However, this method is not applicable for solving sensitive and crucial domains, especially medical data. In fact, the cost of

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

mispredicting minority classes is larger than that of the majority class; this is especially true in medical datasets where high risk patients are the minority class [18]. When compare to other domains, medical dataset is mostly imbalanced [19]. Other common issues in medical data also includes high dimensional data, and it has lower misclassification tolerance [18], [20].

The skewed distribution of multi-class in medical datasets is naturally compounded with many features making them naturally imbalanced [19]. Ideally, classifying medical datasets requires all the classes to retain their form, transforming its initial structure may compromise the validity of diagnosis [21], [22], [23]. The target class of a given medical dataset indicates the severity of the disease for each patient (diagnosis severity class 1,2,3). Medical experts have predetermined the important features to predict the target feature. Decomposing these target classes into a binary class will affect the feature importance during the feature selection process, which leads to bias in the overall predictions [23], [24], [25], [26], [27], [28], [29]. Consequently, it may risk the lives of a patient [19], [21], [23], [28]. Therefore, it is important to retain the classes in their initial form to retain their clinical validity.

There is a need to explore the imbalanced multi-class problem in medical data without decomposition. Studies on this case lack in the body of research [11], [23], [24], [28], [29], [30]. Our previous research [31] reveals that this issue has the most intention in the medical domain. Thus, this current research attempts to address this imbalanced issue. Hence, towards a novel approach, exploring these imbalanced medical data in a rebalancing framework while retaining the multi-class without alteration was thus a major driving force behind this research study.

To the best of our knowledge, few rebalancing frameworks explore this case for medical data. Therefore, we present a new multi-class rebalancing framework using SCUT (SMOTE and Cluster-based Undersampling), RFE (Recursive Feature Elimination), and SHapley Additive exPlanations (SHAP) for feature selection and introduce DES-MI (Dynamic Ensemble Selection for multi-class) for improved multi-classification. The focus of our study is towards rebalancing highly imbalanced datasets. Datasets from the University of California Irvine (UCI), Kaggle, and Knowledge Extraction based on Evolutionary Learning (KEEL) repository were used to validate the proposed rebalancing framework. Furthermore, we also compared the performance of our proposed rebalancing framework with other state-of-the-art imbalanced frameworks and compared the result.

In summary, the key contribution of this article are as follows:

1) In this paper, we introduce a new rebalancing framework for multi-class imbalanced data. A detailed comprehensive analysis of the proposed framework with other state-of-the-art imbalanced frameworks is presented.
2) As a novel approach, we highlight two novelties that contributes to the performance of our framework. Firstly, we improvised SCUT as an improvement by implementing two clustering algorithms, K-means and hierarchical. Secondly, we proposed a pool classifier selection based on extended G-mean(ExGmean) to improve the selection of the candidate pool for DES-MI.
3) Eight imbalanced benchmark datasets are used to validate the framework, with an average of 81.77%, 73.57%, and 75.87% in terms of Macro average accuracy (MAvA), extended G-mean (ExGmean), and Macro Average AUC (MAUC), was attained, respectively. This assures that our proposed framework may also be used on various medical datasets.

The article is organized as follows. Section II reviews the existing related works of literature, and Section III describes the design of the proposed rebalancing framework and more details on the dataset used. Section IV describes the experimental setup, and Section V shows the results of the experiments and discussion. Finally, Section VI concludes the article and discusses the future direction.

II. RELATED WORK

The skewed distribution of classes in medical datasets is naturally compounded with many features. Thus, required an effective rebalancing method that combines feature selection strategies to cater to high dimensionality while maintaining an adequate classification performance.

To accommodate this issue, several endeavours are in the works, particularly a recent work by Krishnan and Sangar [3] that aims to cure the imbalanced nature of medical appointments data in a binary class problem by using different rebalancing techniques unified into one rebalancing framework. Experimental results reveal significant performance. Song et al. [6] proposed a skin cancer melanoma diagnosis that includes a loss function based on focal loss and Jaccard distance to solve the imbalance issue and increase segmentation performances simultaneously. Tested on an imbalanced medical no-show dataset and showed a significant increase in performances. Zhu et al. [32] proposed a hybrid framework that implements an ensemble-based classifier using majority voting with random undersampling. It showed an increase in segmentation performance for binary classification of tumor cancer.

Bi and Ma [7] proposed a similar framework to solve imbalanced cancer datasets for traditional Chinese medicine diagnosis. The framework consists of a three-level structure: data pre-processing, data dimensionality, and rebalancing. In contrast, the first level includes standard data cleaning, while the second level involves the implementation of Long and Short-term Memory Network (LSTM) to reduce the overall dimensionality of the data, and finally, rebalanced using SMOTE. The framework performs significantly well in predicting colorectal cancer. Tang et al. [9] proposed a hybrid framework with a combination of feature selection and ensemble-based learning called the Three-stage Feature selection and Twice-competitional Ensemble learning Method (TSFS-RCEM). This comprises of three-stage; the first stage is to perform information gain (IG) towards the imbalanced data, the second stage involves reducing its high dimensionality, and the final stage, includes feature selection to select the most relevant features.

Sandhan and Choi [8] proposed a framework with an improvised SMOTE that simultaneously rebalances using oversampling and undersampling to prevent the minority class from being neglected during rebalancing. An ensemble-based classifier was used to enhance the classification and proved to significantly increase the overall performance. Likewise, a similar hybrid approach was also studied by Rahim et al. [10] benchmarked on three heart disease datasets. The framework cured the imbalanced issue while performing exceptionally well on cardiovascular disease prediction. Zhao et al. [4] develop a similar rebalancing framework to cater to these imbalanced issues in medical data by using three rebalancing strategies: SMOTE, cost-sensitive learning, over-sampling, and under-sampling technique. The result showed that the rebalancing strategies achieved significant results in imbalanced learning.

While these previous works share a common approach of unifying various rebalancing methods to handle class imbalances, they mostly focus on binary classification problem and uses the decomposition method for such convenience. A direct exploration of the multi-classification problem has not been explored extensively, especially in medical data. Developing a rebalancing framework that caters to multi-class without the need for binary decomposition remains an open challenge. There is also a lack of an adaptive framework that can fulfill both binary and multi-classification issues. Therefore, the goal of this research is to propose a new adaptive framework to overcome this gap.

III. THE PROPOSED MULTI-CLASS REBALANCING FRAMEWORK

In this section, we highlight the overview of the new proposed rebalancing framework and explain each phase. We exhibit which components we adapted and highlight which new components we added to the framework.

A. FRAMEWORK OVERVIEW

Medical dataset has numerous features, and incorporating these attributes is difficult for classification task since it leads to high time complexity and misclassification cost, especially for multi-class. The trade-off between computational cost and the necessity for appropriate class imbalance handling must be carefully considered especially in applications with limited computing capabilities. Therefore, it is essential to choose the appropriate rebalancing strategies that are cost-efficient without compromising computational resources. In this study, we proposed a new rebalancing framework for the multi-classification of imbalanced medical data using multiple combined methods.

The overview of our rebalancing framework is laid out and presented in Figure 1, divided into three phases: phase 1, feature selection and rebalancing; phase 2, training; and phase 3, evaluation and validation. Similar to the basic machine learning lifecycle [33], our framework follows the same workflow. (1) In phase 1, RFE is applied on the imbalance training data and cross reference with SHAP to form the optimal features, then SCUT is used to rebalance the training dataset with the said optimal features, (2) DES-MI is used to train the balanced dataset in phase 2, and (3) finally phase 3, model evaluation using stratified 5-fold cross-validation.

Our proposed framework is entirely new and is an extension to explore the imbalanced issue in medical data for multi-class problems. In actuality, this study is in-line and motivated by similar endeavour work [4]. Therefore, to address the multi-class problem, we highlight the important components that we tune and contributed the most to the performance of our framework: (1) Our novel pool selector by ExGmean for DES-MI. (2) balancing using SCUT with an extension of Kmeans and Hierarchical clustering method. We will explain each component of our framework comprehensively, in the next subsection.

B. PHASE 1: FEATURE SELECTION AND REBALANCING

Medical data is often multi-dimensional, which makes the data mining task much more difficult. Feature selection process is often carried out to mitigate this issue. In our framework, we use a well-known feature selection method, RFE. We then perform a manual cross-reference of the optimal feature with SHAP to further explore the importance and impact of input features. Finally, we perform SCUT to rebalance the dataset. This phase is performed sequentially and the detailed methods are explained below.

1) RECURSIVE FEATURE ELIMINATION

RFE originated from the gene selection research by Guyon et al. [34], since then, it has been efficiently applied in many domains for selecting crucial features [35]. Fundamentally, RFE removes the least important features on the target feature until the optimal features are reached. RFE performs this by the following steps: (1) Computes feature importance using weights to obtain the subset of ranked features. (2) Train the classifier with the subset of features (ranked). (3) Find the feature with the least importance, remove, and update the subset (highest ranking features are eliminated last). (4) Each step is repeated and re-ranks until an optimal ranked list of features (*Foptimal*) with better results is obtained. These procedures are performed iteratively while removing one feature at a time. RFE uses weights (w) to determine the feature importance value. The formula used by RFE to calculate w, for linear problem in Equation (1), where α is the Lagrange coefficients, k is the Kronecker symbol, x is training data, and y is the class label,

$$w = \sum_{k} \alpha_{\mathrm{k}} y_{\mathrm{k}} x_{\mathrm{k}} \tag{1}$$

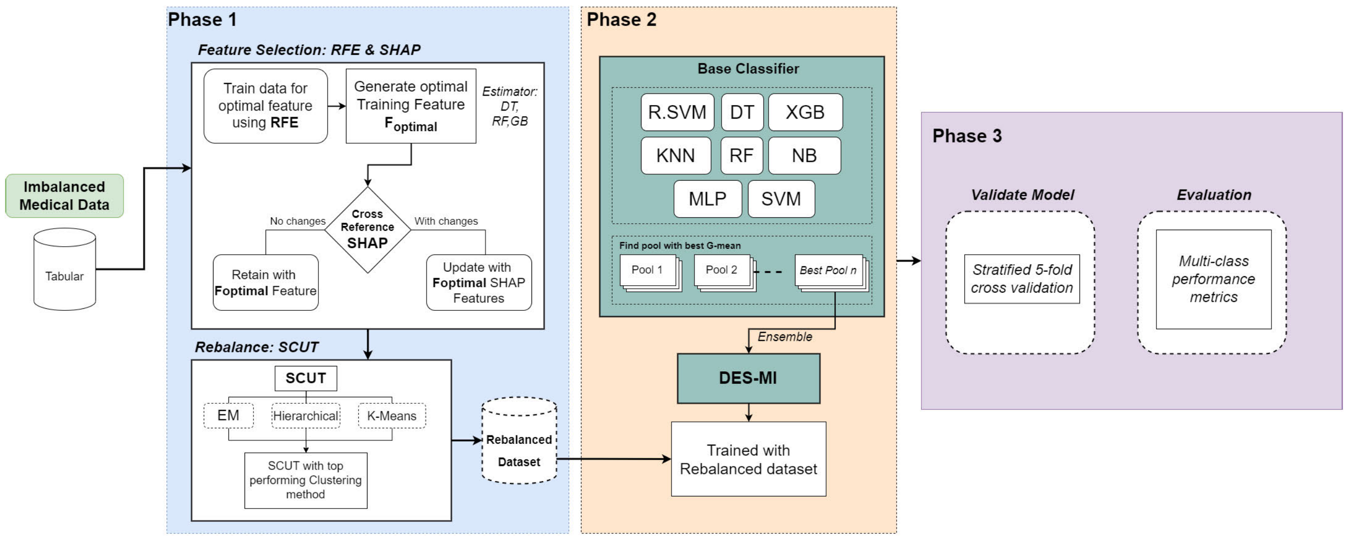


FIGURE 1. **Rebalancing framework for multi-class imbalanced medical data. Foptimal means Optimal features, which are the best-selected features from RFE and SHAP.**

and $DJ(i)$ for the non-linear problem in Equation (2), where i is the feature.

$$DJ(i) = (1/2)(wi)^2 \tag{2}$$

By default, RFE is performed with Support Vector Machine (SVM) classifier [34], however since we experimented on imbalanced data which is mostly non-linear, choosing the appropriate classifier is necessary. Therefore, we decided to experiment on non-linear classifiers (DT, RF, and GB) and obtained better-selected features. The rationale is to choose which classifier provides the highest accuracy based on its selected subset of features (*Foptimal*).

RFE may find important features, that perform best cumulatively in a set. In some sense, these features are optimal when paired together [34], [35]. Since medical datasets are naturally compounded with many features, RFE can be beneficial to find correlated features that might be overlooked in related studies. Additionally, RFE has been proven to be cost-effective [34]. We apply RFE before the rebalancing strategy to retain the feature importance obtained from the initial imbalanced data [36].

2) SHAPLEY ADDITIVE EXPLANATIONS

SHAP is an explainable algorithm popularized by ML researchers to interpret models by demonstrating the impact of each feature on the target class. It is a theoretical approach used to find dominant features by their importance [37]. It has gained popularity for its attribution of interpretable features and has shown to be a reliable alternative feature selection approach [38]. To further dissect the reliability of each feature obtained from the RFE *Foptimal*, we perform a manual cross-reference with SHAP. In detail, SHAP lists out important features ranked by their SHAP value. The higher the value, the higher the priority and its impact on the target class.

The rationale is to cross reference each feature obtained from SHAP and update the *Foptimal* features. For instance, check each feature from both methods and manually remove the feature with the lowest SHAP value, hence update the *Foptimal* features. However, if the list of features from both RFE and SHAP are similar and does not require any changes, no further update is required for *Foptimal*. Subsequently, the imbalance dataset with *Foptimal* features will proceed with the next rebalancing phase.

3) IMPROVED SCUT WITH KMEANS AND HIERARCHICAL CLUSTERING

To cater with the imbalanced medical data issue, we implemented SCUT. A hybridization approach consisting of an oversampling and undersampling method derived from Agrawal et al. [23] specifically to address multi-class imbalance datasets and also applicable for binary class. It oversamples the minority class using SMOTE and then undersamples the majority class using Expectation Maximization (EM) cluster algorithm. EM clustering provides both soft and hard clusters which are already predetermined; thus, it is not necessary to determine the number of clusters in advance.

However, during the experiment, we found out that using EM degrades the overall performance for certain datasets. Apparently, the predetermined clusters do not guarantee an increase in global performance and EM performs slower for larger datasets [39]. To mitigate this limitation, we experimented and compare the results with two well-known clustering algorithms (k-means and hierarchical). We discovered that these two algorithms produced an overall increase in performance and precedes EM in terms of computational cost. Based on this remarkable discovery, we added these algorithms (k-means and hierarchical) as part of SCUT into our proposed framework. We will show the experimental results in a later section.

In **Algorithm 1**, the SCUT with our extension of Kmeans and Hierarchical clustering method is described. SCUT is performed by first splitting the dataset into n parts (target feature), which is $\mathrm{D_i}$ …, $\mathrm{D_n}$, where n is the number of classes and $\mathrm{D_i}$ represents each class. It then calculates the mean (m) of the number of records of all the classes. The algorithm then proceeds with the following conditions: (1) if the number of records (applies in each $\mathrm{D_i}$) is less than mean m, oversampling using SMOTE is performed. The sampling percentage is computed so that the number of instances after oversampling is equal to m. (2) if the number of records (applies in each $\mathrm{D_i}$) is more than the mean m, undersampling is used to generate an equal number of records to the mean m. Our framework improved two clustering algorithms: k-means and hierarchical clustering. The rationale is to choose which algorithm (EM, k-means, or hierarchical) provides optimal results based on the number of clusters it obtains. (3) Else, if the number of records is equal (balance distribution of classes) to the mean m, then SCUT is not performed. Finally, the algorithm proceeds to merge all the classes with an equal number of records to the mean m to produce D', where D' is the balanced dataset.

Additionally, SMOTE performs exceptionally well in reducing class imbalance problems. However, it is inefficient in high-dimensional data, especially in multi-class settings where it is exacerbated the most. Overall, SCUT is far superior to SMOTE for the following two reasons: (1) random sampling weakness, excessive use of both sampling methods may lead to over and underfitting issues. SCUT addresses this issue by finding the correct balance of data while still retaining instances with important information. (2) Accurate for medical data, Agrawal et al. [23] claims that SCUT is appropriate to use for domains that involve multi-class imbalance data and discourage the decomposition method. SCUT tackles this problem by finding the correct balance between class and within-class imbalance by still retaining the multi-class structure.

Additionally, one of the major advantages of SCUT is its capability to handle different types of class imbalance, namely, the between-class and within-class imbalance [23]. The between-class imbalance refers to an imbalance in the distribution of instances across classes and prominently exists in imbalance learning, while the within-class refers to the imbalance that exists within particular target classes which contains variations of underrepresented groups or instances. The rebalancing strategies embedded in SCUT addresses both of these imbalance types, by which SMOTE aids to reduce the between-class issue while the cluster-based undersampling handles the within-class issue.

Algorithm 1 Improved SCUT

Inputs: Dataset D with n classes
Initialize: Divide D into $D_1, D_2, D_3, \ldots, D_\mathrm{n}$, where D_i is a single class, then compute m.
if $D_\mathrm{n} > m$
[Perform Undersampling]
for each D, $i = 1$ to n **do**
  Cluster D_i using EM **or** Kmeans **or** Hierarchical
  for each cluster C_i, $\mathrm{i} = 1, 2, \ldots, k$ **do**
    Randomly select instances from C_i
    Add selected instances to $C_\mathrm{i}{}'$
  end for
  $\mathrm{C} = \emptyset$
  for $i = 1, 2, \ldots, k$ **do**
    $\mathrm{C} = \mathrm{C} \cup \mathrm{C_i}'$
  end for
  $D_\mathrm{i}{}' = \mathrm{C}$
end for
if $D_\mathrm{n} < m$
[Perform SMOTE]
for each D,i=1 to n **do**
  Apply SMOTE on D_i to get $D_\mathrm{i}{}'$
end for
if $D_\mathrm{n} = m$ **then**
  $D_\mathrm{i} = D_\mathrm{i}{}'$
end if
$D' = \emptyset$
for $i = 1$ to n **do**
  $D' = D' \cup D_\mathrm{i}{}'$
end for
return D'
Outputs: Dataset D' has m instances for all classes, where m is the mean instances for all classes.

C. PHASE 2: TRAINING

An obvious distinction between multi-class and its binary counterpart is the decision boundary. Therefore, it is important to find an appropriate classifier that can capture the decision boundary between the majority and minority classes. To solve this issue, we incorporate DES-MI. We also explain our proposed pool classifier selector.

1) DYNAMIC ENSEMBLE SELECTION FOR MULTI-CLASS

DES-MI is a multi-class variant extended from its initial predecessor, the Dynamic Classifier Selection (DCS) and Dynamic Ensemble Selection (DES). On one hand, the DCS method involves selecting a single best classifier for each test sample, on the other, DES involves selecting an optimal classifier ensemble for each test sample. Similar to the former, except its predictions are produced using votes from many classifier models. However, their limitation is only towards binary classification problems. To cater to this restraint, DES-MI was proposed for the multi-class imbalance problem [40]. DES-MI has its perks in solving common multi-class imbalance problem which includes small disjunctions, noise instances, and multi-class overlapping [40]. It solves this by embedding a unique weighting strategy to outperform the competency of a candidate classifier with more strength in identifying the minority classes.

There are two major features of DES-MI summarized as follows: (1) Generation of candidate classifiers. To generate the pool classifier, DES-MI suggests using a homogeneous ensemble (set of classifiers with the same type). In a similar direction to generate the candidate classifiers, we proposed a novel pool classifier selector based on ExGmean, which is shown in **Algorithm 2**. The motivation of the proposed selector is to build an appropriate combination of classifiers in the pool rather than the manual selection which is time-consuming. In this way, an adequate diversity of top-performed classifiers by ExGmean can be formed. The experiment of our proposed pool selector will be shown in the later section. (2) Dynamic selection of the most appropriate ensemble. Introduces the novel weighting strategy to outperform the competency of a candidate classifier (in pool classifier) with more strength in identifying minority classes. In other words, higher weights were given when measuring the competency level of a classifier in the candidate classifier pool.

In **Algorithm 3**, the detailed DES-MI algorithm is described. The algorithm proceeds as follows: (1) evaluate the performance of potential classifiers in their region of competence for each query sample that is required to be classified, denote as X_t, and X_i evaluates the impact of the class. The region of competence is defined by the k nearest neighbors around the query example. (2) The algorithm's main purpose is to pick classifiers that are stronger when categorizing cases that are underrepresented in the region of competence. Each classifier's competence (in the classifier pool) is determined, and the adaptive weight adjustment process is applied. Classifiers with more strength in categorizing complicated instances from all classes are associated with better competency. (3) The selected classifiers are combined decisively using a majority vote.

2) PROPOSED POOL CLASSIFIER SELECTION BASED ON EXTENDED G-MEAN

While still maintaining the diversity, our approach of selecting the classifiers for the pool focuses on classifiers that provide the best ExGmean score. The most dependable metric that reflects the overall model performance across all classes is the g-mean. Therefore, it is only reasonable to include classifiers that provide the best g-mean into the pool. However, the conventional g-mean metric is built for binary classification, to extend it for multi-class, we decided to use the extended G-mean instead [29]. ExGmean is calculated by Equation (3).

$$ExGmean = (\sum_{i=1}^{k} R_i)^{1/k} \tag{3}$$

where R is the recall and k is the class.

The candidate base classifiers that we use for the pool are c4.5 Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGB), Support Vector Machine (SVM), Radial kernel Support Vector Machine (R.SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN), and Multilayer Perceptron (MLP). DT algorithm is a popular approach for classification and is extensively used to solve medical diagnoses [30]. Meanwhile, RF can avoid complexity issues due to its immunity towards overfitting and the curse of dimensionality [41]. XGB is an extended version of a gradient boosting approach that produces a robust boosted tree model with good accuracy and is well-known for its resistance to imbalanced data [42]. Both SVM and R.SVM have reputable classification performance in the medical area for disease prediction [43]. NB and MLB are well-known for imbalanced learning [44], [45]. While KNN has been used by many for its reliability in disease prediction [46]. Since this study includes imbalanced medical data, it is reasonable to include these classifiers as the base candidate for our pool classifier.

The procedure of our proposed pool classifier selection based on the ExGmean method is described in the pseudocode in **Algorithm 2**. Given a dataset, D, split into the training set, D_{tr}, and testing set D_{te}, with candidate classifier pool, denote as CP. For each loop, train and predict each classifier CP_i using the ExGmean algorithm (Equation (3)) and store it in a new finalized classifier pool, CP_{gmean}. To generate the top N performed classifier, simply update the n_{top}. By default, we set the n_{top} equal to CP_n. The rationale for choosing the best n_{top} classifier from CP_{gmean} is to check which classifiers in the pool work best with each other as pairs or individuals. It is not always a good idea to use every possible classifier in the CP_{gmean}, as certain datasets may work best and others may degrade. It is best to preserve a wide range of candidate classifiers while limiting the ones with low g-mean. Therefore, any n_{top} variants of the CP_{gmean} can be produced for experimentation. For instance, variant 1, CP_{gmean} (n_{top}=6) with six highest g-mean classifiers from the pool; variant 3, CP_{gmean}(n_{top}=3) with only the top three highest g-mean classifiers.

Algorithm 2 Pseudocode for Proposed Pool Classifier Selector by ExGmean

Inputs: Training set D_{tr}, testing set D_{te}, candidate classifiers pool CP, n_{top} best-performed classifiers by gmean value
$D' \leftarrow \emptyset$
for each CP, $i = 1$ to CP_n **do**
  Train CP_i on D_{tr}
  scores $\leftarrow$ store CP_i gmean score
  predict CP_i on D_{te} using ExGmean algorithm
  $CP_{gmean} = (CP_i, \text{scores})$
end for
Sort CP_{gmean} by ascending value
$CP_{gmean} = CP_{gmean}$ (n_{top})
Outputs: CP_{gmean}, proposed classifier pool with best gmean

Algorithm 3 DES-MI

Inputs: Training set D_{tr}, validation set D_{valid}, testing set D_{te}, nearest neighbors k, percentage of classifiers to be selected P%, the scaling coefficient α, and our proposed classifier pool CP_{gmean}
for each X_{t} in D_{te} **do**
  $EoC*_{\mathrm{t}}' \leftarrow \emptyset$
  find Ψ as the k nearest neighbours of the instances X_{t} in D_{valid}
  for each X_{i} in Ψ **do**
    num $\leftarrow$ **count** number of instances with the same class as X_{i}
    $W_{\mathrm{i}} \leftarrow 1/1+\exp(\alpha \times num)$ //calculate the voting weights for each X_{i} in Ψ
  end for
  Normalize W_{i} according to $\hat{W}_{\mathrm{i}} \leftarrow \frac{W_{\mathrm{i}}}{\sum_{i=1}^{k} W_{\mathrm{i}}}$
  for each classifier CP_{j} in CP_{gmean} **do**
    $C(CP_{\mathrm{j}}|X_{\mathrm{t}}) \leftarrow \sum_{i=1}^{k} I(CP_{\mathrm{j}}(X_{\mathrm{t}}) = y_{\mathrm{t}})\hat{W}_{\mathrm{i}}$
  end for
  select P% most competent classifiers in CP_{j} to create the ensemble $EoC*_{\mathrm{t}}$ for instances X_{t}
  $H(X_{\mathrm{t}}) \leftarrow \arg\max_{y \in \Omega} \sum_{i=1}^{N} I(h_{\mathrm{i}}(X_{\mathrm{t}}) = y)$
end for
Outputs: CP_{gmean}, proposed classifier pool with best gmean

D. PHASE 3: EVALUATION AND VALIDATION

It is imperative to choose appropriate evaluation metrics to measure the multi-class. Accuracy is not a valid metric in imbalanced data with multi-class settings. Therefore, instead of using the standard metrics (recall, accuracy, precision, f-score), we will use Macro average accuracy (MAvA), ExGmean, and Macro Average AUC (MAUC), to properly evaluate the overall performances across all the classes. These alternative measures are most suited to measure each class's performance in a multi-class problem [11], [47].

Due to the class instances in some of the datasets (eColi, Yeast, Lymphography) being relatively small. We use a stratified 5-fold cross-validation to validate the model and ensure each minority class has at least one example in each fold and is appropriate for imbalance learning [28], [48]. Additionally, previous related works [3], [9], [10] have shown that using a 5-fold cv approach provides significant results. Furthermore, we used the default parameter and manually performed the cross-validation without the *"pipeline"* python library. This is by means to make the framework computationally cost-efficient.

IV. EXPERIMENTAL VERIFICATION AND SETUP

To verify the performance of our proposed rebalancing framework we compare it in two aspects; (1) with no framework applied, we depict this as, *Standard* approach, and (2) with other state-of-the-art imbalanced learning frameworks. Table 1 describes the details of the different state-of-the-art imbalance frameworks for comparison. While these frameworks share distinct approaches in solving class imbalances, they are limited to only binary class.

TABLE 1. Details of the different state-of-the-art imbalance frameworks for comparison.

Author	Year	Classifiers	Feature Selection	Rebalancing Strategies
Krishnan & Sangkar [3]	2021	DT	N/A	RUS,ROS,SMOTE, ENN, and CNN
Rahim et al. [10]	2021	Boosting with KNN and LR	Feature Importance	SMOTE
Błaszczyk & Jedrzejowicz [49]	2021	RF,KNN,DT,MLP, NB, and SVM	N/A	KNORA-E,KNORA-U, and KNORA-P.
Bashir et al. [50]	2016	QDA, LR, NB, KNN, SVM, DT(Info gain and Gini Index)	F1 Feature Selection	MMV

LR=Logistic Regression, QDA=Quadratic Discriminant Analysis, RUS=Random Undersampling, ROS=Random Oversampling, ENN=Edited Nearest Neighbor, KNORA-E=K-Nearest-Oracle-Eliminate,KNORA-U=K-Nearest-Oracle-Union,KNORA-P=K-Nearest-Oracle-Performance, MMV=Multilayer Majority Voting, N/A=Not Available

The overall experimental workflow is shown in Figure 2. The experimental results were obtained using stratified 5-fold cross-validation with 3 iterations. Each dataset was divided into five folds, with each fold holding 20% of the dataset's instances as the test set and the remaining 80% as the training set.

A. DATASETS

Table 2 shows the eight imbalanced medical datasets used for the experiment. The UCI datasets are eColi and Yeast. Cirrhosis, HepatitisC, Framingham, Stroke, and MIMIC-III were obtained from Kaggle and Lymphography are obtained from KEEL. For this study, all the datasets have varying levels of imbalance and multiple classes. To make our framework more generalizable for binary class problem we include a dataset that has binary class hence, Framingham, Stroke, and MIMIC-III was included.

Imbalance Ratio (IR) represents the level of imbalance on each dataset. The smaller the IR, the more balanced the dataset; hence, the distribution will be less skewed. However, the larger the IR, the larger the imbalanced extent of the dataset [5]. A mild IR is between 1.9 and 9, while highly IR is more than 9 [51]. In this case, there are five highly imbalanced dataset with IR more than 9. Yeast dataset has the highest level of imbalance with IR=92.6 followed by eColi, Lymphography, HepatitisC, and Stroke datasets with IR=71.92, 40.5, 25.71, and 19.52, respectively. The IR is calculated by dividing the highest-class ratio, $r_{\max}$ by the lowest-valued class ratio, $r_{\min}$. Measuring the IR helps us identify the imbalance level on each dataset we used. The IR is calculated by Equation (4).

$$IR = \frac{r_{\mathrm{maj}}}{r_{\min}} \tag{4}$$

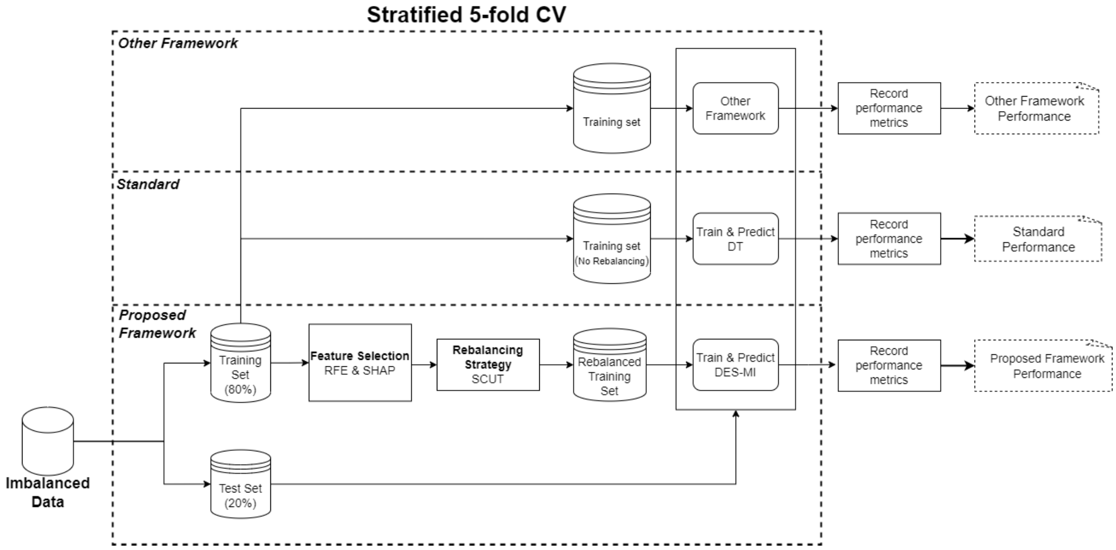


FIGURE 2. Experimental workflow.

B. PERFORMANCE EVALUATION METRICS FOR MULTI-CLASS CLASSIFICATION

In a binary class setting, the standard metric commonly used to measure the performance of a model are accuracy, precision, recall, and f-score. By definition, accuracy represents an overall predictive capability of a model [47], [52]. Each class has the same weight which contributes equally to the overall accuracy. Precision denotes the proportion of values that a model predicts will be positive (TP, true positive). It indicates how much we can rely on the model when the predictive value is positive [47]. Recall, assesses the model's prediction accuracy for the positive class. Which is calculated by dividing the proportion of TP values by the total number of positive values [47], [52]. The f-score evaluates a classification model performance by averaging both precision and recall measurements using the harmonic mean approach [47]. The higher the f1-score the better the predictive capabilities of each class. However, we evaluate the performance of our model using three main evaluation metrics for multi-class; these metrics are MAvA, ExGmean, and MAUC.

ExGmean is used to measure the mean recall of all the classes due to its sensitivity towards the minority class [29] and efficiency to identify the minority class. This metric is defined in Equation (3).

MAvA is a more effective metric that measures the average accuracy across all classes in a multi-class setting. Defined as the arithmetic mean of each class's partial accuracies [53].The formula for MAvA is calculated in Equation (5).

$$MAvA = \frac{\sum_{i=1}^{J} ACC_{\mathrm{i}}}{J} \tag{5}$$

where J is a class.

The Receiver Operating Characteristic curve (ROC) is one of the most extensively used tools for evaluating binary classifiers for imbalanced learning. It describes the trade-off between the true positive rate (TPR) and the false positive rate (FPR) for a predictive model with varying probability thresholds. The area under the ROC is called AUC. To adapt the AUC for multi-class problems, we will use the MAUC. It computes and averages the AUC of each class vs the rest (one vs. rest) [54]. The MAUC is calculated in Equation (6).

$$MAUC = \frac{1}{J}\sum_{j \in J} AUC_{\mathrm{R}}(j, rest_{\mathrm{j}}) \tag{6}$$

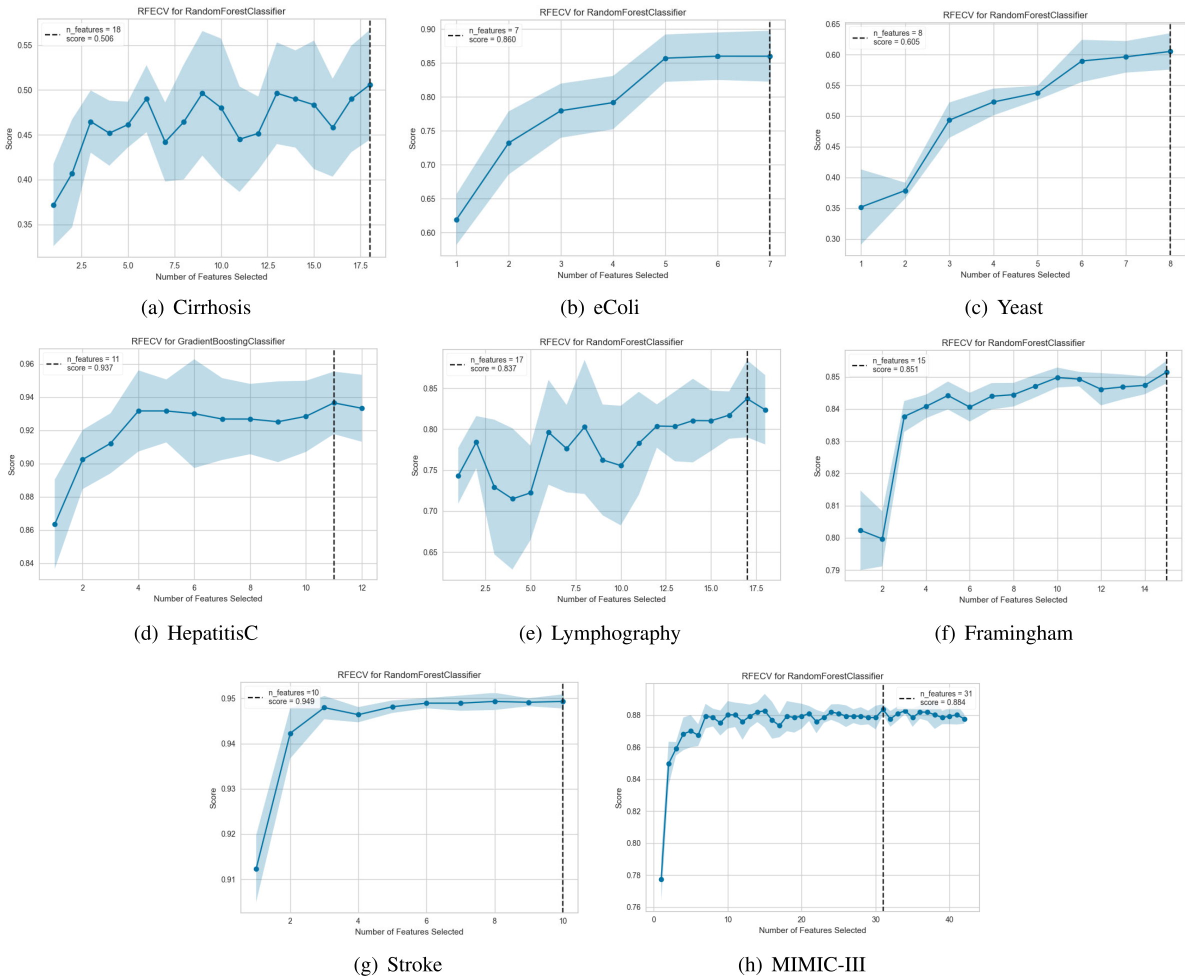
C. STATISTICAL TEST

Statistical techniques must be used to analyze the results to determine whether there are significant differences between the proposed framework and the other state-of-the-art framework. Therefore, the Wilcoxon signed-rank test [55] was used in this case for the pairwise comparison as a non-parametric hypothesis test. In details, rankings "1" and "0" are given for the outcomes of the two methods with the lowest and highest absolute disparities. To calculate R^+ and R^-, the ranks of positive and negative differences are added. Whereas, the p-value in Wilcoxon signed-rank test [55] shows a significant difference between the two methods if it is less than the significance level of 0.05.

V. EXPERIMENT RESULTS AND DISCUSSION

A. FEATURE SELECTION WITH RFE AND SHAP

The RFE was performed with stratified 5-fold cross-validation on each fold and was averaged to determine the *Foptimal* features by accuracy. The following parameter was used for the cross-validation: *n_split* = 5, *shuffle* = True, *random_state* = 42, and *scoring='accuracy'*. Figure 3 shows



(a) Cirrhosis (b) eColi (c) Yeast

(d) HepatitisC (e) Lymphography (f) Framingham

(g) Stroke (h) MIMIC-III

FIGURE 3. Selected *Foptimal* features using RFE on each dataset.

the selected *Foptimal* by the number of features using RFE on each dataset. Next, we perform a SHAP analysis to compute the SHAP value of each feature towards the target class. A higher SHAP value indicates the impact of the feature on the model's output. We then perform a manual cross reference on the RFE *Foptimal* with SHAP to analyze features with the lowest impact and remove them. For instance, in the Lymphography dataset, feature *bl_of_lymph_s* has the lowest rank on RFE, and the lowest SHAP value, since it is the lowest feature on both sides, we remove the feature and update the *Foptimal* list. Note that we only remove the lowest features that are correspondence on both RFE and SHAP (highlighted in bold). Table 3 shows the summary of RFE and SHAP feature selection.

B. SCUT AND OPTIMAL CLUSTERING METHOD

We performed a stratified 5-fold cross-validation, 80% of the training data was rebalanced with SCUT of each clustering method. Trained and tested on the remaining 20% set. The following parameter was used for the cross-validation: *n_split* = *5*, *shuffle* = *True*, and *random_state* = *7*. We compare the results to determine which method produces the best ExGmean. Table 4 shows the performance of three clustering types for optimal clustering method on each dataset by ExGmean. Note that, we only choose the highest ExGmean (highlighted in bold) as the optimal clustering for each dataset.

According to Table 4, the hierarchical method provides a higher ExGmean for most of the dataset with HepatitisC, Lymphography, Framingham, Stroke, and MIMIC-III. While eColi and Yeast performed well with EM, K-means algorithm only perform best for Cirrhosis. The results also showed that the execution time(s) for EM was 245.87s, slower than that of K-means with 228.35s, and Hierarchical being the fastest with 213.28s. This approves the limitations of EM [39] being computationally expensive for larger datasets (Stroke, MIMIC, Framingham) and degradation of results.

C. DES-MI WITH BEST ExGmean POOL CLASSIFIER

DES-MI requires a pool of classifiers to work properly with any number of candidate classifiers. In our study, there are eight base candidate classifiers in the pool, *CP*. To highlight

TABLE 2. Summary of imbalanced medical datasets.

Benchmark Dataset	No. of Records	No. of Features	Class Distribution	Class Ratios, r	IR, r_{max}/r_{min}	Imb. Type
Cirrhosis	418	18	Class=0, n=21	0.050*	7.67	WC
			Class=1, n=92	0.22		
			Class=2, n=161	0.3852*		
			Class=3, n=144	0.344		
HepatitisC	615	12	Class=0, n=540	0.8780*	25.71	WC
			Class=1, n=24	0.0390		
			Class=2, n=21	0.0341*		
			Class=3, n=30	0.0488		
eColi	337	7	Class=0, n=143	0.4243*	71.92	BC
			Class=1, n=77	0.2285		
			Class=2, n=2	0.0059		
			Class=3, n=2	0.0059*		
			Class=4, n=35	0.1038		
			Class=5, n=20	0.059		
			Class=6, n=5	0.01485		
			Class=7, n=52	0.1543		
Yeast	1484	8	Class=0, n=463	0.3120*	92.60	BC
			Class=1, n=429	0.2891		
			Class=2, n=244	0.1644		
			Class=3, n=163	0.1098		
			Class=4, n=51	0.0344		
			Class=5, n=44	0.0296		
			Class=6, n=35	0.0236		
			Class=7, n=30	0.0202		
			Class=8, n=20	0.0135		
			Class=9, n=5	0.0034*		
Lymphography	148	18	Class=0, n=2	0.0135	40.5	WC
			Class=1, n=81	0.5473*		
			Class=2, n=61	0.4122		
			Class=3, n=4	0.0270*		
Framingham	4133	15	Class=0, n=3505	0.8481*	5.58	BC
			Class=1, n=628	0.1519*		
Stroke	5110	10	Class=0, n=4861	0.9513*	19.52	BC
			Class=1, n=249	0.0487*		
MIMIC-III	1175	48	Class=0, n=1016	0.8647*	6.39	BC
			Class=1, n=159	0.1353*		

Imb. Type=Imbalance Type, WC=Within-Class, BC=Between-Class

TABLE 3. Summary of feature selection using RFE and SHAP.

Dataset	RFE			SHAP
	Acc	Foptimal	Lowest Rank	Lowest IF
Cirrhosis	0.50	18	None	None
eColi	0.86	7	None	None
Yeast	0.60	8	None	None
HepatitisC	0.94	11	**Sex**	**Sex**
Lymphography	0.84	17	**bl_of_lymph_s**	**bl_of_lymph_s**
Framingham	0.85	15	None	None
Stroke	0.95	10	None	None
MIMIC-III	0.88	31	**age,MCHC,EF, deficiencyanemias, hypertensive, atrialfibrillation, gendera, hyperlipemia, diabetes, depression,**	**age,atrialfibrillation, Renalfailure,EF deficiencyanemias, MCHC, hypertensive hyperlipemia, gendera, diabetes, depression**

Acc=Accuracy, IF=Impact Factor

the strength of our proposed selector we compared it with the standard pool with no selector. We implement our pool classifier selector by ExGmean in **Algorithm 2** to get the CP_{gmean}. To find the best n_{top} we created two different pools and compared them with the standard pool. The details of the pools are as follows:

TABLE 4. ExGmean performance of three clustering types for optimal clustering method on each dataset.

Dataset	EM	Hierarchical	K-Means
Cirrhosis	0.5891	0.5881	**0.6161**
eColi	**0.8909**	0.8791	0.8747
Yeast	**0.660**	0.6538	0.6456
HepatitisC	0.7286	**0.8035**	0.7184
Lymphography	0.8569	**0.8599**	0.85345
Framingham	0.4932	**0.6109**	0.4447
Stroke	0.5440	**0.5980**	0.5487
MIMIC-III	0.6480	**0.6914**	0.6407
Execution Time(s)	245.87	**213.28**	228.35

TABLE 5. Selected CP_{gmean} and n_{top} for each dataset by ExGmean.

Dataset	Pool 1	Pool 2	Pool 3
Cirrhosis	0.6271	0.59	**0.6288**
eColi	**0.9268**	0.8881	0.9014
Yeast	0.6704	0.619	**0.6875**
HepatitisC	0.7652	0.7218	**0.8130**
Lymphography	0.8055	0.8048	**0.9546**
Framingham	0.5511	0.567	**0.6109**
Stroke	**0.6919**	0.6846	0.5624
MIMIC-III	0.6761	0.6082	**0.6767**

- Pool 1: Standard, a pool with all the candidate classifier and no selector.
- Pool 2: $CP_{\text{gmean}}(n_{\text{top}}=6)$, a pool with the top 6 candidate classifier.
- Pool 3: $CP_{\text{gmean}}(n_{\text{top}}=4)$, a pool with the top 4 candidate classifier.

We train and test each dataset and choose the ones that provide the best results. Table 5 shows the best pool for each dataset (highlighted in bold). Based on the results, most of the datasets performed well with Pool 3 and only two datasets (eColi and Stroke) performed best with Pool 1. In most cases, the dataset provided better results with the top 4 classifiers.

D. PERFORMANCE COMPARISON VERSUS STANDARD APPROACH AND OTHER STATE-OF-THE-ART IMBALANCED FRAMEWORK

To compare the effectiveness of our proposed framework we compare it with *Standard* and other state-of-the-art imbalanced frameworks. We performed stratified 5-fold cross-validation with three iterations under the following parameter: *n-split=5* and *shuffle=True*. The random_state (seed) for each iteration are 7, 24, and 42 to make the experiment study more reproducible for interested readers. The hyperparameter was set to default for all models. We obtained the results and recorded them as average validation performance across 5 folds. Table 6 shows the average 5-fold of each dataset for *Standard* and other state-of-the-art imbalanced frameworks. The mean summary of ExGmean and MAUC are reported in Figures 4 and 5. The ExGmean and MAUC performance of our proposed framework by each iteration is depicted in Figure 6. Additionally, we have performed the required Wilcoxon signed-ranked test for the pairwise comparison between the proposed framework with *Standard* and other state-of-the-art frameworks. The statistical results in terms of ExGmean and MAUC are shown in Table 7, where the ranks for the proposed framework and the compared frameworks are added up to form R^+ and R^-, respectively.

The analysis based on the experimental results is as follows:

1) According to Table 6, the results highlighted in bold demonstrated the best overall performance on each dataset, in terms of MAvA, ExGmean, and MAUC. The *Standard* approach achieves higher MAvA but has relatively low ExGmean and MAUC across all the datasets. The imbalanced data distribution adds to this degradation. In a sense that it will cause many misclassifications of predictive outcomes, hence, the lower ExGmean and MAUC. Meanwhile, all the other state-of-the-art imbalanced frameworks performed better than the *Standard* approach with adequate ExGmean and MAUC. Evidently, our proposed framework obtained significant MAvA, ExGmean, and MAUC across all datasets. Except for Stroke and MIMIC with the Standard approach has the highest MAvA with a trade-off of lower ExGmean and MAUC.
2) As presented in Figures 4 and 5, our proposed framework achieves the best robustness by mean ExGmean of 5.87 and MAUC of 6.07 compared to the other state-of-the-art frameworks, across all the datasets. The *Standard* approach with no rebalancing achieved the lowest ExGmean while DT+CNN achieved the lowest MAUC. Evidently, the model′s performances across all the dataset remained fairly consistent across each iteration, indicating no signs of overfitting.
3) According to Figure 6, the results shows that our framework achieved a consistent ExGmean(a) and MAUC(b) across all the dataset especially eColi with above 0.90 ExGmean and MAUC. Most of the dataset has consistent results by iterations.
4) Referring to Table 6, our proposed framework performs significantly better than most of the frameworks due to the corresponding p-values being less than the significant value of 0.05. Although there are no significant differences(p-values>0.05) found between proposed vs. KNORAE+ROS and proposed vs. KNORAE+SMOTE in terms of ExGmean and MAUC, we can, however, emphasize the strong performance of the proposed framework since in both instances, the values of R^+ are significantly higher than those of R^-. Also, for ExGmean and MAUC, our proposed framework achieved the most wins with 8 out of 8 datasets by the majority of the pairwise framework comparisons. However, our proposed framework wins 4 out of 8, and 5 out of 8 datasets when compared to KNORAE+ROS and KNORAE+SMOTE in terms of ExGmean. Whereas, in terms of MAUC, our proposed framework wins by 5 out of 8 datasets for both KNORAE+ROS and KNORAE+SMOTE, respectively.

Table 8 presents the overall average results of our proposed framework with *Standard* and State-of-the-art Imbalanced Framework. The best result is highlighted in bold. According to Table 8, our proposed framework achieved the highest ExGmean and MAUC overall by 0.7357 and 0.7587, respectively. Apparently, the *Standard* approach has the highest MAvA with 0.8517 among the others. However, it suffers from lower ExGmean and MAUC. Signifying that model with no framework produced lower predictive performance on the minority class. Evidently, the results of our proposed framework outperform the other State-of-the-art Imbalanced Frameworks and *Standard* approaches with a significant increase in overall metrics especially for ExGmean and MAUC. Although each state-of-the-art framework performs with equivalent results, it does make a clear distinction when comparing the results with our proposed framework.

E. DISCUSSION

It can be observed from the results that our proposed framework outperforms the *Standard* approach and other state-of-the-art imbalanced frameworks with significant overall performance across all the imbalanced datasets. Specifically in terms of the multi-class metrics, MAvA, ExGmean, and MAUC. Results from these experiments conclude that our proposed framework significantly solves the imbalance data issue and improves the overall performances while retaining the multi-class setting in medical data and without the decomposition method. It clarifies the limitations of the other state-of-the-art imbalance framework which is limited to only binary class, our framework was able to solve both binary and multi-class settings.

These results are consistent with previous related works using a rebalancing framework to solve the imbalance distribution of classes [3], [6], [10] in medical datasets and notably prior similar works that incorporate various rebalancing strategies and feature selection unified into one framework [7], [9]. Previous findings that implement an ensemble-based classifier [8], [32] as part of their rebalancing framework also show similar results in solving these imbalanced datasets. The significance and contributions of this study are summarised below based on the results of the experiment:

1) Handling imbalanced issues in medical data: This study proposed a rebalancing framework to solve the class imbalanced issue that resides in medical data. Thus, based on the results, the improved SCUT can handle the imbalanced class issue with the best MAvA, ExGmean, and MAUC. Additionally, the feature selection method using RFE and SHAP can reduce data dimensionality and increase sensitivity, hence the increase in ExGmean. The applied DES-MI can reduce classification error and improve classification performance by giving weights to each class.
2) Highly imbalanced ratio: Our proposed rebalancing framework can solve highly imbalanced medical datasets notably datasets with IR of more than 9.

TABLE 6. Comparison of our proposed framework with standard and state-of-the-art imbalanced framework for each imbalanced datasets.

Dataset	Reference	Framework	Results		
			MAvA	ExGmean	MAUC
Lymphography	Krishnan & Sangkar [3]	DT+RUS	0.6025	0.6098	0.6978
		DT+ROS	0.7620	0.7654	0.8010
		DT+SMOTE	0.7627	0.7658	0.8042
		DT+ENN	0.5201	0.5073	0.6687
		DT+CNN	0.5047	0.5165	0.6478
	Rahim et al. [10]	Boosting+FI+SMOTE	0.7574	0.7535	0.8202
	Błaszczyk & Jedrzejowicz [49]	KNORAE+ROS	0.7876	0.7835	0.8458
		KNORAE+SMOTE	0.8480	0.8456	0.8822
		KNORAU+ROS	0.8018	0.7980	0.8595
		KNORAU+SMOTE	0.8569	0.8543	0.8907
		KNORAP+ROS	0.7941	0.7905	0.8521
		KNORAP+SMOTE	0.8358	0.8339	0.8705
	Bashir et al. [50]	MLV+FS	0.6947	0.6519	0.7383
	Standard	No Framework	0.8355	0.5864	0.7160
	Our Proposed	**Our Framework**	**0.9278**	**0.8954**	**0.9135**
eColi	Krishnan & Sangkar [3]	DT+RUS	0.2886	0.3345	0.5339
		DT+ROS	0.6382	0.6658	0.7597
		DT+SMOTE	0.6328	0.6987	0.7996
		DT+ENN	0.6464	0.7095	0.7890
		DT+CNN	0.4079	0.4824	0.6653
	Rahim et al. [10]	Boosting+FI+SMOTE	0.6854	0.6676	0.7429
	Błaszczyk & Jedrzejowicz [49]	KNORAE+ROS	0.7137	0.7227	0.8014
		KNORAE+SMOTE	0.7158	0.7360	0.8001
		KNORAU+ROS	0.7115	0.7027	0.7767
		KNORAU+SMOTE	0.7443	0.7323	0.8378
		KNORAP+ROS	0.7105	0.7302	0.8179
		KNORAP+SMOTE	0.7304	0.7829	0.8298
	Bashir et al. [50]	MLV+FS	0.6406	0.7030	0.8032
	Standard	No Framework	0.9447	0.5627	0.6678
	Our Proposed	**Our Framework**	**0.9598**	**0.9207**	**0.9265**
Cirrhosis	Krishnan & Sangkar [3]	DT+RUS	0.3627	0.5113	0.5743
		DT+ROS	0.3803	0.5152	0.5807
		DT+SMOTE	0.2916	0.4209	0.5311
		DT+ENN	0.3679	0.4917	0.5723
		DT+CNN	0.3139	0.4529	0.5434
	Rahim et al. [10]	Boosting+FI+SMOTE	0.4157	0.5840	0.6345
	Błaszczyk & Jedrzejowicz [49]	KNORAE+ROS	0.4028	0.5239	0.6086
		KNORAE+SMOTE	0.4176	0.5701	0.6130
		KNORAU+ROS	0.4170	0.5339	0.6168
		KNORAU+SMOTE	0.4341	0.5804	0.6254
		KNORAP+ROS	0.4118	0.5072	0.6142
		KNORAP+SMOTE	0.4190	0.5534	0.6168
	Bashir et al. [50]	MLV+FS	0.3663	0.5132	0.5793
	Standard	No Framework	0.6974	0.4771	0.5641
	Our Proposed	**Our Framework**	**0.7326**	**0.6197**	**0.6568**
HepatitisC	Krishnan & Sangkar [3]	DT+RUS	0.6208	0.7249	0.7728
		DT+ROS	0.5248	0.5933	0.7106
		DT+SMOTE	0.5469	0.6589	0.7378
		DT+ENN	0.5405	0.5924	0.7264
		DT+CNN	0.5642	0.6816	0.7234
	Rahim et al. [10]	Boosting+FI+SMOTE	0.5875	0.7080	0.7452
	Błaszczyk & Jedrzejowicz [49]	KNORAE+ROS	0.6069	0.7012	0.7844
		KNORAE+SMOTE	0.6915	0.7897	0.8209
		KNORAU+ROS	0.6597	0.7620	0.8012
		KNORAU+SMOTE	0.6141	0.7215	0.7724
		KNORAP+ROS	0.6453	0.7358	0.7944
		KNORAP+SMOTE	0.6582	0.7522	0.8036
	Bashir et al. [50]	MLV+FS	0.5943	0.6990	0.7539
	Standard	No Framework	0.9528	0.6269	0.7239
	Our Proposed	**Our Framework**	**0.9130**	**0.7960**	**0.8239**

TABLE 6. *(Continued.)* Comparison of our proposed framework with standard and state-of-the-art imbalanced framework for each imbalanced datasets.

Yeast	Krishnan & Sangkar [3]	DT+RUS	0.3774	0.5230	0.6468
		DT+ROS	0.4657	0.5944	0.7004
		DT+SMOTE	0.4796	0.6144	0.7082
		DT+ENN	0.4528	0.5663	0.6973
		DT+CNN	0.2364	0.3631	0.5718
	Rahim et al. [10]	Boosting+FI+SMOTE	0.5568	0.6850	0.7488
	Błaszczyk & Jedrzejowicz [49]	KNORAE+ROS	0.4603	0.5852	0.6985
		KNORAE+SMOTE	0.5289	0.6664	0.7310
		KNORAU+ROS	0.4688	0.5885	0.7050
		KNORAU+SMOTE	0.5218	0.6617	0.7314
		KNORAP+ROS	0.5148	0.6404	0.7265
		KNORAP+SMOTE	0.4907	0.6180	0.7148
	Bashir et al. [50]	MLV+FS	0.4587	0.5711	0.7000
	Standard	No Framework	0.9030	0.4990	0.6584
	Our Proposed	**Our Framework**	**0.9067**	**0.6833**	**0.7513**
Framingham	Krishnan & Sangkar [3]	DT+RUS	0.5738	0.5735	0.5738
		DT+ROS	0.5358	0.4319	0.5358
		DT+SMOTE	0.5379	0.4941	0.5379
		DT+ENN	0.5793	0.5470	0.5793
		DT+CNN	0.5502	0.5340	0.5502
	Rahim et al. [10]	Boosting+FI+SMOTE	0.5616	0.4962	0.5616
	Błaszczyk & Jedrzejowicz [49]	KNORAE+ROS	0.5333	0.3568	0.5333
		KNORAE+SMOTE	0.5771	0.5638	0.5771
		KNORAU+ROS	0.5411	0.3979	0.5411
		KNORAU+SMOTE	0.5730	0.5478	0.5730
		KNORAP+ROS	0.5674	0.4518	0.5674
		KNORAP+SMOTE	0.5679	0.5124	0.5679
	Bashir et al. [50]	MLV+FS	0.5370	0.3377	0.5370
	Standard	No Framework	0.7605	0.4669	0.5536
	Our Proposed	**Our Framework**	**0.6362**	**0.5913**	**0.6019**
Stroke	Krishnan & Sangkar [3]	DT+RUS	0.6492	0.6381	0.6492
		DT+ROS	0.5500	0.3643	0.5500
		DT+SMOTE	0.5618	0.4476	0.5618
		DT+ENN	0.5922	0.4902	0.5922
		DT+CNN	0.5604	0.4786	0.5604
	Rahim et al. [10]	Boosting+FI+SMOTE	0.6383	0.5869	0.6383
	Błaszczyk & Jedrzejowicz [49]	KNORAE+ROS	0.5609	0.3883	0.5609
		KNORAE+SMOTE	0.6032	0.5382	0.6032
		KNORAU+ROS	0.6168	0.5342	0.6168
		KNORAU+SMOTE	0.6378	0.5990	0.6378
		KNORAP+ROS	0.5961	0.4948	0.5961
		KNORAP+SMOTE	0.6398	0.6039	0.6398
	Bashir et al. [50]	MLV+FS	0.5250	0.2332	0.5250
	Standard	No Framework	**0.9090**	0.3500	0.5406
	Our Proposed	**Our Framework**	0.6775	**0.6875**	**0.6895**
MIMIC	Krishnan & Sangkar [3]	DT+RUS	0.6369	0.6356	0.6369
		DT+ROS	0.5565	0.4550	0.5565
		DT+SMOTE	0.6160	0.5830	0.6160
		DT+ENN	0.6327	0.5997	0.6327
		DT+CNN	0.6168	0.6078	0.6168
	Rahim et al. [10]	Boosting+FI+SMOTE	0.6671	0.6303	0.6671
	Błaszczyk & Jedrzejowicz [49]	KNORAE+ROS	0.5937	0.4476	0.5937
		KNORAE+SMOTE	0.6332	0.5568	0.6333
		KNORAU+ROS	0.6303	0.5223	0.6303
		KNORAU+SMOTE	0.6749	0.6277	0.6749
		KNORAP+ROS	0.6520	0.5770	0.6520
		KNORAP+SMOTE	0.6468	0.5732	0.6468
	Bashir et al. [50]	MLV+FS	0.6006	0.4652	0.6006
	Standard	No Framework	**0.8111**	0.5336	0.6070
	Our Proposed	**Our Framework**	0.7882	**0.6914**	**0.7060**

We experimented on eight imbalanced datasets, each with varying levels of IR, five of which are highly imbalanced (Yeast, eColi, Lymphography, HepatitisC, and Stroke). We include three binary-class datasets to

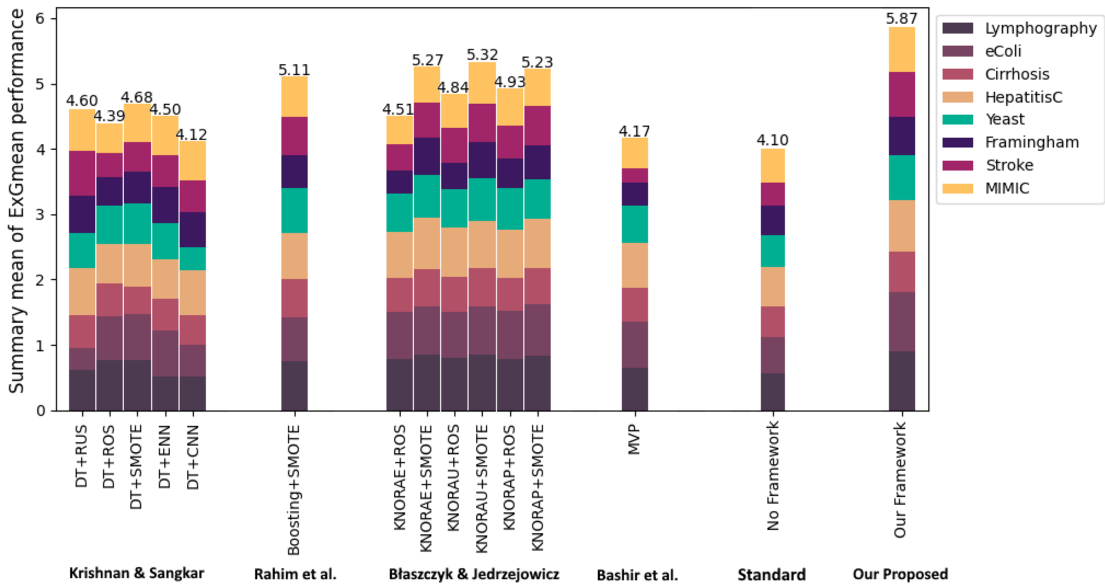


FIGURE 4. Summary mean performance of each framework in terms of ExGmean.

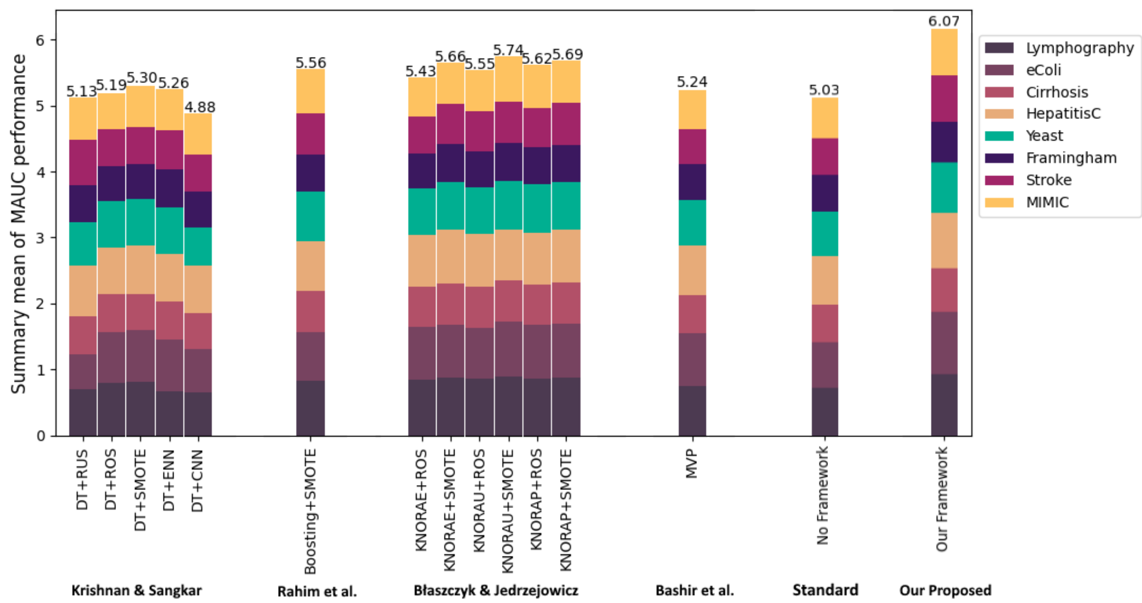
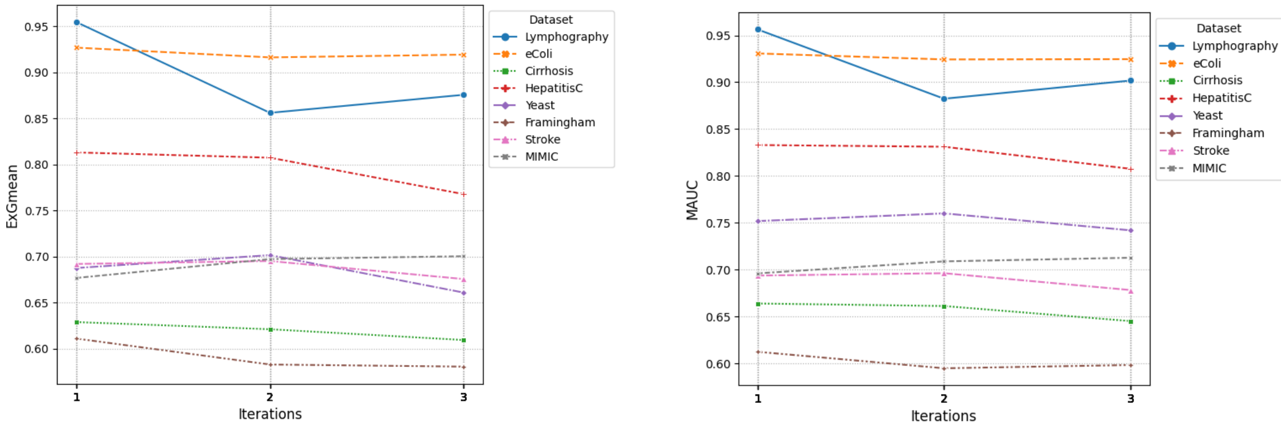


FIGURE 5. Summary mean performance of each framework in terms of MAUC.

increase the generality of our framework not limited to only multi-class problems. Unfortunately, we could not acquire more medical dataset that is highly imbalanced in nature. Despite this limitation, our framework was able to solve three highly imbalanced dataset, thus, provide ample sufficiency of our framework's capability that contribute towards its robustness in solving highly imbalanced dataset.

3) Novelties that contribute to the performance of our framework: (1) We implement SCUT as part of the rebalancing strategy in our proposed framework, it uses EM as its standard clustering algorithm. However, our experiment (using SCUT with EM) showed degrades in overall performance for certain datasets and an increase in time complexity. We experimented and compared the results with k-means and hierarchical. The results show an overall increase in performance and thus, precedes the limitations of EM. Thus, this provision indicates an improvement in SCUT. (2) We introduced our pool classifier selector by ExGmean, as an appropriate selector of candidate classifier for DES-MI. Our approach effortlessly finds the suitable pool classifier



(a) ExGmean

(b) MAUC

FIGURE 6. ExGmean performance of proposed framework by iterations.

TABLE 7. Wilcoxon's test for pairwise comparison between our proposed framework and state-of-the-art framework by ExGmean and MAUC.

Metrics	Framework Comparison	Results				
		W/L	R^+	R^-	P-value	Significant (Yes/No)
ExGmean	Proposed vs. DT+RUS	8/0	120.00	0	0.043	Yes
	Proposed vs. DT+ROS	8/0	120.00	0	0.043	Yes
	Proposed vs. DT+SMOTE	8/0	120.00	0	0.043	Yes
	Proposed vs. DT+ENN	8/0	120.00	0	0.043	Yes
	Proposed vs. DT+CNN	8/0	120.00	0	0.043	Yes
	Proposed vs. Boosting+FI+SMOTE	8/0	120.00	0	0.043	Yes
	Proposed vs. KNORAE+ROS	4/4	94.00	26.00	0.07025	No
	Proposed vs. KNORAE+SMOTE	5/3	79.00	41.00	0.064125	No
	Proposed vs. KNORAU+ROS	7/1	106.00	14.00	0.0476	Yes
	Proposed vs. KNORAU+SMOTE	7/1	107.00	13.00	0.049	Yes
	Proposed vs. KNORAP+ROS	7/1	111.00	9.00	0.0462	Yes
	Proposed vs. KNORAP+SMOTE	8/0	120.00	0	0.043	Yes
	Proposed vs. MLV+FS	8/0	120.00	0	0.043	Yes
	Proposed vs. Standard	8/0	120.00	0	0.043	Yes
MAUC	Proposed vs. DT+RUS	8/0	120.00	0	0.043	Yes
	Proposed vs. DT+ROS	8/0	120.00	0	0.043	Yes
	Proposed vs. DT+SMOTE	8/0	120.00	0	0.043	Yes
	Proposed vs. DT+ENN	8/0	120.00	0	0.043	Yes
	Proposed vs. DT+CNN	8/0	120.00	0	0.043	Yes
	Proposed vs. Boosting+FI+SMOTE	8/0	120.00	0	0.043	Yes
	Proposed vs. KNORAE+ROS	5/3	81.00	39.00	0.09	No
	Proposed vs. KNORAE+SMOTE	5/3	80.00	40.00	0.0714	No
	Proposed vs. KNORAU+ROS	7/1	106.00	14.00	0.0476	Yes
	Proposed vs. KNORAU+SMOTE	7/1	108.00	12.00	0.0486	Yes
	Proposed vs. KNORAP+ROS	7/1	111.00	9.00	0.0462	Yes
	Proposed vs. KNORAP+SMOTE	8/0	120.00	0	0.043	Yes
	Proposed vs. MLV+FS	8/0	120.00	0	0.043	Yes
	Proposed vs. Standard	8/0	120.00	0	0.043	Yes

for DES-MI rather than choosing the classifier manually. We hope that this novel method may assist interested researchers in imbalanced learning when using DES methods. In fact, our novel pool selector can also be implemented with other DES variants not limited to DES-MI, as long as that particular variant supports the pool classifier parameter.

Notably, increasing the generalizability of the framework in handling both classification tasks (binary and multi-class) comes with a trade-off of limiting its practicality, especially in real-world applications where efficiency and responsiveness are critical. For that reason, hyperparameter tuning and grid search are not included in this framework, particularly in SCUT and DES-MI methods. As clarity, the purpose of the enhanced SCUT with two additional clustering algorithms is by means to address the time consumption issue in EM with improved results. While the rest of the SCUT was performed in the default parameter setting with no further optimization tuning. The same can also be said for the proposed pool classifier for DES-MI, where its purpose is only to obtain the suitable candidate classifiers ranked by ExGmean. While proceeding with the rest of the algorithm remains unhinged.

However, it is essential to evaluate the proposed framework on more new data to ensure their performance remains

TABLE 8. Overall average results of our proposed framework with standard and state-of-the-art imbalanced framework.

Author	Framework	Results		
		MAvA	ExGmean	MAUC
Krishnan & Sangkar [3]	DT+RUS	0.5140	0.5688	0.6357
	DT+ROS	0.5517	0.5482	0.6493
	DT+SMOTE	0.5537	0.5854	0.6621
	DT+ENN	0.5415	0.5630	0.6572
	DT+CNN	0.4693	0.5146	0.6099
Rahim et al. [10]	Boosting+FI+SMOTE	0.6087	0.6087	0.6087
Błaszczyk & Jedrzejowicz [49]	KNORAE+ROS	0.5824	0.5637	0.6783
	KNORAE+SMOTE	0.6269	0.6583	0.7076
	KNORAU+ROS	0.6059	0.6049	0.6934
	KNORAU+SMOTE	0.6321	0.6656	0.7179
	KNORAP+ROS	0.6115	0.6160	0.7026
	KNORAP+SMOTE	0.6236	0.6537	0.7113
Bashi et al. [50]	MLV+FS	0.6087	0.6087	0.6087
Standard	No Framework	**0.8517**	0.5128	0.6289
Our Proposed	**Our Framework**	0.8177	**0.7357**	**0.7587**

consistent over time, given that data distribution changes constantly. Unfortunately, acquiring more imbalanced medical data has become challenging due to its scarcity and data-sharing restrictions. Regardless, in future works, we intended to validate and refine the framework based on more new real-world data to ensure its practicality and generalizability in a dynamic and evolving environment.

Overall, this study highlights the capability of our framework in solving multi-class imbalanced medical data, leading to effective rebalancing and an increase in overall performance. Furthermore, the statistical analysis using Wilcoxon signed-rank test for pairwise comparison shows that our proposed framework significantly outperforms the *Standard* and the other state-of-the-art frameworks. However, it is worth noting that our framework is not limited to medical data; but is also, applicable to rebalance datasets that have similar unbalanced distribution in different data domains as well.

VI. CONCLUSION AND FUTURE DIRECTION

Class imbalance exists in many data domains, especially for medical datasets, which are inevitably imbalanced in nature. For a more convenient solution, most researchers preferred the standard method of decomposing multi-classes into subproblems of binary classes. This approach, however, is not applicable for the sensitive and critical domain, likewise, medical data. The fact that clinical validity requires, all classes to preserve their shape to avoid the diagnosis from being compromised.

In this work presented, we present a new multi-class rebalancing framework using SCUT, RFE, and SHAP for feature selection, and introduce DES-MI with our novel pool selector by ExGmean, for improved multi-classification. This rebalancing framework was experimented with using eight imbalanced medical datasets UCI, Kaggle, and KEEL repositories. Experiments were carried out, and results showed that our proposed rebalancing framework demonstrates a significant overall performance that outperforms the *Standard* approach and other state-of-the-art imbalanced frameworks. In terms of multi-class performance metrics MAvA, ExGmean, and MAUC.

In the hope of validating and further improving our rebalancing framework, it is of our research interest to experiment with more highly imbalanced datasets and explore other common medical data issues such as high dimensionality and misclassification tolerance. As a future intention, we plan to not only explore real-world imbalanced datasets from diverse domains beyond medical but also to delve into the impact of hyperparameters within our proposed framework. This exploration aims to uncover the sensitivity of the introduced techniques to hyperparameter variations, ultimately guiding the selection of optimal parameters for improved results.

CONFLICT OF INTEREST

The authors have no conflict of interest to mention.

REFERENCES

[1] S. M. A. Elrahman and A. Abraham, "A review of class imbalance problem," *J. Netw. Innov. Comput.*, vol. 1, no. 8, pp. 332–340, 2013.

[2] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*, vol. 21. Cham, Switzerland: Springer, 2018.

[3] U. Krishnan and P. Sangar, "A rebalancing framework for classification of imbalanced medical appointment no-show data," *J. Data Inf. Sci.*, vol. 6, no. 1, pp. 178–192, Feb. 2021, doi: 10.2478/jdis-2021-0011.

[4] Y. Zhao, Z. S. Y. Wong, and K. L. Tsui, "A framework of rebalancing imbalanced healthcare data for rare events' classification: A case of look-alike sound-alike mix-up incident detection," *J. Healthcare Eng.*, vol. 2018, May 2010, Art. no. 6275435, doi: 10.1155/2018/6275435.

[5] R. Zhu, Y. Guo, and J.-H. Xue, "Adjusting the imbalance ratio by the dimensionality of imbalanced data," *Pattern Recognit. Lett.*, vol. 133, pp. 217–223, May 2020, doi: 10.1016/j.patrec.2020.03.004.

[6] L. Song, J. Lin, Z. J. Wang, and H. Wang, "An end-to-end multi-task deep learning framework for skin lesion analysis," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 2912–2921, Oct. 2020, doi: 10.1109/JBHI.2020.2973614.

[7] W. Bi and R. Ma, "Unbalanced data set processing method for colorectal cancer prediction in TCM diagnosis," in *Proc. IEEE Int. Conf. E-Health Netw., Appl. Services (HEALTHCOM)*, Mar. 2021, pp. 1–6, doi: 10.1109/HEALTHCOM49281.2021.9615914.

[8] T. Sandhan and J. Y. Choi, "Handling imbalanced datasets by partially guided hybrid sampling for pattern recognition," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 1449–1453, doi: 10.1109/ICPR.2014.258.

[9] X. Tang, L. Cai, Y. Meng, C. Gu, J. Yang, and J. Yang, "A novel hybrid feature selection and ensemble learning framework for unbalanced cancer data diagnosis with transcriptome and functional proteomic," *IEEE Access*, vol. 9, pp. 51659–51668, 2021, doi: 10.1109/ACCESS.2021.3070428.

[10] A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim, and A. W. Muzaffar, "An integrated machine learning framework for effective prediction of cardiovascular diseases," *IEEE Access*, vol. 9, pp. 106575–106588, 2021, doi: 10.1109/ACCESS.2021.3098688.

[11] P. Phoungphol, Y. Zhang, and Y. Zhao, "Robust multiclass classification for learning from imbalanced biomedical data," *Tsinghua Sci. Technol.*, vol. 17, no. 6, pp. 619–628, Dec. 2012, doi: 10.1109/TST.2012.6374363.

[12] K. Hee-Sung, "Impact of web-based nurse's education on glycosylated haemoglobin in type 2 diabetic patients," *J. Clin. Nurs.*, vol. 16, no. 7, pp. 1361–1366, Apr. 2007, doi: 10.1111/j.1365-2702.2007.01506.x.

[13] R. J. Cascaro, B. D. Gerardo, and R. P. Medina, "Filter selection methods for multiclass classification," in *Proc. ACM 2nd Int. Conf. Comput. Big Data*, 2019, pp. 27–31, doi: 10.1145/3366650.3366655.

[14] Hartono, S. Lestari, A. Rahmadsyah, R. M. F. Lubis, and M. Gunawan, "HAR-MI with COSTE in handling multi-class imbalance," in *Proc. 8th Int. Conf. Cyber IT Serv. Manag. (CITSM)*, Oct. 2020, pp. 16–19, doi: 10.1109/CITSM50537.2020.9268804.

[15] N. K. Singha Roy and B. Rossi, "Cost-sensitive strategies for data imbalance in bug severity classification: Experimental results," in *Proc. 43rd Euromicro Conf. Softw. Eng. Adv. Appl. (SEAA)*, Aug. 2017, pp. 426–429, doi: 10.1109/SEAA.2017.71.

[16] T. Wang, K.-C. Shu, C.-H. Chang, and Y.-F. Chen, "On the effect of data imbalance for multi-label pedestrian attribute recognition," in *Proc. Conf. Technol. Appl. Artif. Intell. (TAAI)*, Nov. 2018, pp. 74–77, doi: 10.1109/TAAI.2018.00025.

[17] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS Int. Trans. Comput. Sci. Eng.*, vol. 30, no. 1, pp. 25–36, 2006. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.96.9248&rep=rep1&type=pdf

[18] N. Liu, X. Li, E. Qi, M. Xu, L. Li, and B. Gao, "A novel ensemble learning paradigm for medical diagnosis with imbalanced data," *IEEE Access*, vol. 8, pp. 171263–171280, 2020, doi: 10.1109/ACCESS.2020.3014362.

[19] B. M. Bai, N. Mangathayaru, and B. P. Rani, "Exploring research issues in mining medical datasets," in *Proc. Int. Conf. Eng. MIS (ICEMIS)*, vol. 24, Sep. 2015, pp. 1–8, doi: 10.1145/2832987.2833078.

[20] M. M. Rahman and D. N. Davis, "Addressing the class imbalance problem in medical datasets," *Int. J. Mach. Learn. Comput.*, vol. 3, pp. 224–228, Apr. 2013, doi: 10.7763/ijmlc.2013.v3.307.

[21] N. Salkind, *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA, USA: Sage Publications, 2007.

[22] Y. Hu and M. Sokolova, "Explainable multi-class classification of medical data," 2020, *arXiv:2012.13796*.

[23] A. Agrawal, H. L. Viktor, and E. Paquet, "SCUT: Multi-class imbalanced data classification using SMOTE and cluster-based undersampling," in *Proc. 7th Int. Joint Conf. Knowl. Discov. Knowl. Eng. Knowl. Manag. (IC3K)*, vol. 1, Nov. 2015, pp. 226–234, doi: 10.5220/0005595502260234.

[24] M. Bouazizi and T. Ohtsuki, "Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in Twitter," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6, doi: 10.1109/ICC.2016.7511392.

[25] S. Okada, M. Ohzeki, and S. Taguchi, "Efficient partition of integer optimization problems with one-hot encoding," *Sci. Rep.*, vol. 9, no. 1, pp. 1–12, Sep. 2019, doi: 10.1038/s41598-019-49539-6.

[26] A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care," *Neural Comput. Appl.*, vol. 32, no. 24, pp. 18069–18083, Dec. 2020, doi: 10.1007/s00521-019-04051-w.

[27] O. Niaksu, "CRISP data mining methodology extension for medical domain," *Baltic J. Mod. Comput.*, vol. 3, no. 2, pp. 92–109, 2015.

[28] S. Wang and X. Yao, "Multiclass imbalance problems: Analysis and potential solutions," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 4, pp. 1119–1130, Aug. 2012, doi: 10.1109/TSMCB.2012.2187280.

[29] Y. Sun, M. S. Kamel, and Y. Wang, "Boosting for learning multiple classes with imbalanced class distribution," in *Proc. 6th Int. Conf. Data Mining (ICDM)*, Dec. 2006, pp. 592–602, doi: 10.1109/ICDM.2006.29.

[30] J. D. Pineda-Jaramillo, "A review of Machine Learning (ML) algorithms used for modeling travel mode choice," *DYNA*, vol. 86, no. 211, pp. 32–41, Oct. 2019, doi: 10.15446/dyna.v86n211.79743.

[31] J. Edward and M. M. Rosli, "A systematic mapping study on ensemble-based classifier," in *Proc. IEEE Int. Conf. Comput. (ICOCO)*, Nov. 2021, pp. 43–48, doi: 10.1109/ICOCO53166.2021.9673563.

[32] W. Zhu, B.-S. Oh, W. Huang, Z. Lin, Y. Pan, and J. Zhou, "Hybrid classifiers ensemble with an undersampling scheme for liver tumor segmentation," in *Proc. 10th Int. Conf. Inf., Commun. Signal Process. (ICICS)*, Dec. 2015, pp. 1–4, doi: 10.1109/ICICS.2015.7459850.

[33] O. Spjuth, J. Frid, and A. Hellander, "The machine learning life cycle and the cloud: Implications for drug discovery," *Expert Opinion Drug Discovery*, vol. 16, no. 9, pp. 1071–1079, Sep. 2021, doi: 10.1080/17460441.2021.1932812.

[34] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, Jan. 2002, doi: 10.1023/A:1012487302797.

[35] X.-W. Chen and J. C. Jeong, "Enhanced recursive feature elimination," in *Proc. 6th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2007, pp. 429–435, doi: 10.1109/ICMLA.2007.35.

[36] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinf.*, vol. 14, no. 1, p. 106, Dec. 2013, doi: 10.1186/1471-2105-14-106.

[37] R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu, S. Xu, S. Barb, A. Joseph, M. Shumski, J. Smith, A. B. Sood, G. S. Corrado, L. Peng, and D. R. Webster, "Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy," *Ophthalmology*, vol. 126, no. 4, pp. 552–564, Apr. 2019, doi: 10.1016/j.ophtha.2018.11.016.

[38] D. Fryer, I. Strümke, and H. Nguyen, "Shapley values for feature selection: The good, the bad, and the axioms," *IEEE Access*, vol. 9, pp. 144352–144360, 2021, doi: 10.1109/ACCESS.2021.3119110.

[39] Y. G. Jung, M. S. Kang, and J. Heo, "Clustering performance comparison using K-means and expectation maximization algorithms," *Biotechnol. Biotechnol. Equip.*, vol. 28, no. sup1, pp. S44–S48, Nov. 2014, doi: 10.1080/13102818.2014.949045.

[40] S. García, Z.-L. Zhang, A. Altalhi, S. Alshomrani, and F. Herrera, "Dynamic ensemble selection for multi-class imbalanced datasets," *Inf. Sci.*, vols. 445–446, pp. 22–37, Jun. 2018, doi: 10.1016/j.ins.2018.03.002.

[41] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[42] N. Cahyana, S. Khomsah, and A. S. Aribowo, "Improving imbalanced dataset classification using oversampling and gradient boosting," in *Proc. 5th Int. Conf. Sci. Inf. Technol. (ICSITech)*, Oct. 2019, pp. 217–222, doi: 10.1109/ICSITech46713.2019.8987499.

[43] K. Harimoorthy and M. Thangavelu, "Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 3, pp. 3715–3723, Mar. 2021, doi: 10.1007/s12652-019-01652-0.

[44] R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel, "Constrained Naïve Bayes with application to unbalanced data classification," *Central Eur. J. Oper. Res.*, vol. 30, no. 4, pp. 1403–1425, Dec. 2022, doi: 10.1007/s10100-021-00782-1.

[45] Z. Zhao, S. Xu, B. H. Kang, M. M. J. Kabir, Y. Liu, and R. Wasinger, "Investigation and improvement of multi-layer perceptron neural networks for credit scoring," *Expert Syst. Appl.*, vol. 42, no. 7, pp. 3508–3516, May 2015, doi: 10.1016/j.eswa.2014.12.006.

[46] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of heart disease using K-nearest neighbor and genetic algorithm," *Proc. Technol.*, vol. 10, pp. 85–94, Dec. 2013, doi: 10.1016/j.protcy.2013.12.340.

[47] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: An overview," 2020, *arXiv:2008.05756*.

[48] V. López, A. Fernández, and F. Herrera, "On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed," *Inf. Sci.*, vol. 257, pp. 1–13, Feb. 2014, doi: 10.1016/j.ins.2013.09.038.

[49] M. Błaszczyk and J. Jędrzejowicz, "Framework for imbalanced data classification," *Proc. Comput. Sci.*, vol. 192, pp. 3477–3486, Jan. 2021, doi: 10.1016/j.procs.2021.09.121.

[50] S. Bashir, U. Qamar, F. H. Khan, and L. Naseem, "HMV: A medical decision support framework using multi-layer classifiers for disease prediction," *J. Comput. Sci.*, vol. 13, pp. 10–25, Mar. 2016, doi: 10.1016/j.jocs.2016.01.001.
[51] N. Noorhalim, A. Ali, and S. M. Shamsuddin, "Handling imbalanced ratio for class imbalance problem using SMOTE," in *Proc. 3rd Int. Conf. Comput., Math. Statist. (iCMS)*. Singapore: Springer, Mar. 2019, pp. 19–30.
[52] R. A. Sowah, B. Kuditchar, G. A. Mills, A. Acakpovi, R. A. Twum, G. Buah, and R. Agboyi, "HCBST: An efficient hybrid sampling technique for class imbalance problems," *ACM Trans. Knowl. Discovery Data*, vol. 16, no. 3, pp. 1–37, Jun. 2022, doi: 10.1145/3488280.
[53] J. P. Sánchez-Crisostomo, R. Alejo, E. López-González, R. M. Valdovinos, and J. H. Pacheco-Sánchez, "Empirical analysis of assessments metrics for multi-class imbalance learning on the back-propagation context," in *Advances in Swarm Intelligence* (Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8795. Cham, Switzerland: Springer, Oct. 2014, pp. 17–23.
[54] C. Ferri, J. Hernández-Orallo, and M. A. Salido, "Volume under the ROC surface for multi-class problems," in *Machine Learning: ECML 2003* (Lecture Notes in Artificial Intelligence and Lecture Notes in Computer Science), vol. 2837. Berlin, Germany: Springer, 2003, pp. 108–120, doi: 10.1007/978-3-540-39857-8_12.
[55] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, Jul. 1945, doi: 10.2307/3001968.

JAFHATE EDWARD received the B.S. degree in computer science and the M.S. degree in computer education from Universiti Malaysia Sabah (UMS), Malaysia, in 2016 and 2020, respectively. He is currently pursuing the Ph.D. degree in computer science with the MARA University of Technology (UiTM), Malaysia. His research interests include artificial intelligence and machine learning in the applications of medical domain.

MARSHIMA MOHD ROSLI received the B.Sc. degree (Hons.) in information technology from Universiti Utara Malaysia, in 2001, the M.Sc. degree in real-time software engineering from Universiti Teknologi Malaysia, in 2006, and the Ph.D. degree in computer science from the University of Auckland, New Zealand, in 2018. She is currently an Associate Professor with the Department of Computer Science, College of Computing, Informatics and Mathematics, MARA University of Technology (UiTM), Malaysia, where she has been a Faculty Member, since 2007. Her research interests are primarily in the area of software engineering, artificial intelligent, and data analytics.

ALI SEMAN received the Ph.D. degree in computer science/bioinformatics from the MARA University of Technology (UiTM), Malaysia, in 2013. He has 26 years working experience, including eight years in industry and 18 years in academic. His research interest includes algorithms and optimization, particularly in the area of clustering and classification.