Pertanika J. Sci. & Technol. 32 (6): 2631 - 2653 (2024)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

A Comprehensive Analysis of a Framework for Rebalancing Imbalanced Medical Data Using an Ensemble-based Classifier

Jafhate Edward[1], Marshima Mohd Rosli[1,2]* and Ali Seman[1]

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

[2]Institute for Pathology, Laboratory and Forensic Medicine (I-PPerForM), Universiti Teknologi MARA, 47000 UiTM, Sungai Buloh, Selangor, Malaysia

ABSTRACT

In medical data, addressing imbalanced datasets is paramount for accurate predictive modeling. This paper delves into exploring a well-established rebalancing framework proposed in previous research. While acknowledged for its effectiveness, the adaptability of this framework across diverse medical datasets remains unexplored. We conduct a comprehensive investigation to bridge this gap by integrating an ensemble-based classifier into the existing framework. By leveraging seven imbalanced medical binary datasets, our study comprises three distinct experiments: utilizing standard baseline classifiers from the framework (original), incorporating the baseline with an ensemble-based classifier, and introducing our novel ensemble-based classifier with the self-paced ensemble (SPE) algorithm. Our novel ensemble, composed of decision tree (DT), radial support vector machine (R.SVM), and extreme gradient boosting (XGB) classifiers, serves as the foundation for the SPE. Our primary objective is to demonstrate the potential improvement of the existing framework's overall performance through the integration of an ensemble. Experimental results reveal significant enhancements, with our proposed ensemble classifier outperforming the original by 4.96%, 5.89%, 5.68%, 7.85%, and 6.84% in terms of accuracy, precision, recall, F-score, and G-mean, respectively. This study contributes valuable insights into the adaptability and performance augmentation achievable through ensemble methods in addressing class imbalances within the medical domain.

Keywords: Ensemble classifier, ensemble learning, imbalance classification, machine learning algorithms, medical data, predictive modeling, rebalancing framework

ARTICLE INFO

Article history:
Received: 14 December 2023
Accepted: 10 Jun 2024
Published: 25 October 2024

DOI: https://doi.org/10.47836/pjst.32.6.12

E-mail addresses:
jafhate@gmail.com (Jafhate Edward)
marshima@uitm.edu.my (Marshima Mohd Rosli)
aliseman@uitm.edu.my (Ali Seman)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

Class imbalance poses a common challenge across various data domains, particularly in medical datasets where its presence is unavoidable (Bai et al., 2015; Rahman & Davis, 2013). Imbalance occurs when the majority class overwhelms instances of the minority class (Abraham & Elrahman, 2013). Predictive models trained on imbalanced data often exhibit bias, resulting in a higher misclassification rate when predicting the target outcome. Conventional sampling methods, such as random oversampling, undersampling, and the synthetic minority oversampling technique (SMOTE), are commonly applied to address this issue. These methods involve modifying imbalanced datasets to create a more balanced distribution, significantly improving the overall performance of classifiers (Fernández et al., 2018).

In the medical domain, the consequences of such misclassifications can be considerably more significant, as they may lead to the misdiagnosis of cancerous patients as noncancerous or vice versa (Belarouci & Chikh, 2017; Rahman & Davis, 2013). Consequently, various cutting-edge techniques for dealing with this issue have emerged (Rahman & Davis, 2013; Pes, 2019). One such approach employs the rebalancing framework proposed by Zhao et al. (2018), which implements several standard classifiers and relies on four rebalancing strategies to address data imbalance. A satisfactory increase in overall classification performance has been reported by Zhao et al. (2018), especially in terms of accuracy, recall, precision, and F-score.

Despite promising results, the framework's effectiveness with medical datasets of varying imbalance levels remains unexplored. This uncharted nature motivates the current research study. However, the validity and effectiveness of the rebalancing framework may vary between datasets based on the imbalance ratio (IR), necessitating classifiers capable of enhancing performance across different IR (Mohammed et al., 2020; Krishnan & Sangar, 2021; Tantithamthavorn et al., 2020; Jiang et al., 2020).

Therefore, our research investigates the effectiveness of ensemble-based classifiers within the Zhao et al. (2018) rebalancing framework while exploring additional datasets, aligning with the author's attention to experimentally test the framework with more imbalanced medical datasets.

An ensemble-based classifier is a combination of more than one classifier (Valentini & Dietterich, 2004) that performs better than individual ones. Researchers across various domains have widely implemented ensembles for enhanced classification (Mohandes et al., 2018). The advantages of employing an ensemble approach are: (1) it combines stronger classifiers to address class imbalance, ensuring efficient imbalance learning (Khalilia et al., 2011; Cahyana et al., 2019); (2) our previous work (Edward & Rosli, 2021), a systematic mapping study (SMS) on ensemble-based classifiers, highlighted the favorability of the ensemble approach among researchers in the medical domain, particularly for its

effectiveness in diagnostic classification. The SMS also uncovered a prevailing trend where many researchers prefer the hard-level majority voting technique as their primary choice for ensemble combination methods, especially in medical research. Thus, we selected the hard-level majority voting technique as our preferred combination method. These reasons alone were enough to spark our interest in exploring this method via Zhao et al. (2018) rebalancing framework.

To further leverage the capabilities of the assembling approach, we adapted our proposed ensemble-based classifier as the baseline estimator within the Self-paced Ensemble (SPE), an imbalance learning method introduced by Liu et al. (2020). SPE introduces the classification 'hardness' concept to demonstrate a trained classifier's difficulty in identifying a particular sample. Based on this hardness, SPE iteratively selects the most informative majority data samples in accordance with their distribution rather than simply balancing the positive and negative data or applying instance weights. Implementing SPE for highly imbalanced data is expected to yield significant results (Liu et al., 2020).

This study conducts three experiments to determine whether ensemble-based classifiers improve the existing framework. Initially, we assessed the performance of each imbalanced dataset used in this study. We compared the results obtained using the original baseline classifier recommended by Zhao et al. (2018) framework with those achieved with ensemble-based baseline classifiers. Subsequently, we conducted another experiment with our proposed ensemble-based classifier to evaluate the effectiveness of the ensemble approach. The results were then comprehensively compared to determine which method demonstrated a more substantial performance improvement.

The overall results of our experiments revealed that our proposed ensemble-based classifiers with SPE outperformed both the original baseline and the baseline with the ensemble approach in terms of overall performance measures (accuracy, precision, recall, F-score, and G-mean). This outcome highlights the effectiveness of the ensemble method in addressing class imbalances in medical data, demonstrating its potential for enhanced performance in imbalance learning. In summary, the key contributions of this article are as follows:

1. To investigate and provide a comprehensive analysis of the effectiveness of ensemble-based classifiers in the rebalancing framework proposed by Zhao et al. (2018).
2. To explore and experimentally test the framework of Zhao et al. (2018) with more imbalanced medical datasets.
3. To introduce and evaluate SPE(EM), a novel ensemble approach. SPE(EM), combining decision tree (DT), radial support vector machine (R.SVM), and extreme gradient boosting (XGB) classifiers, outperformed the baseline with significant improvements (4.96%, 5.89%, 5.68%, 7.85%, and 6.84%) in accuracy,

precision, recall, F-score, and G-mean. This contribution extends the understanding of ensemble methods in addressing class imbalances in medical datasets.

THE FRAMEWORK

As previously mentioned, the framework we adopted and tested in this research study is the rebalancing framework developed by Zhao et al. (2018). In their research, the authors experimented with medical incidents due to look-alike (LASA) mix-ups dataset, which exhibited class imbalance. This dataset comprises 227 records with structured text, including eight features and binary class target variables (LASA and non-LASA). The authors' framework demonstrated a notable ability to classify LASA incident reports with high predictive accuracy. Although their primary focus was on incident report classification, the authors suggested that their rebalancing framework holds broad applicability, extending beyond the classification of medical incident reports to address other medical datasets with similar imbalanced properties.

Zhao et al.'s (2018) framework incorporates algorithmic and data-level approaches to rebalance the unequal class distribution to address the class imbalance issue. A detailed investigation was conducted to assess the impact and performance of various classifiers, utilizing a sequence of three key stages within the framework. Specifically, these stages are based on classifier selection, incorporating four rebalancing strategies, and leave-one-out cross-validation (LOOCV). In the initial stage, the performance of each candidate classifier is evaluated based on standard metrics (accuracy, precision, recall, and F-score) to determine the best-performing classifier. Zhao et al. (2018) suggest that candidate classifiers can be linear or non-linear for binary classification. However, for their experimental studies, they opted for logistic regression (LR), support vector machine with linear kernels (L.SVM), support vector machine with radial kernels (R.SVM), and decision tree (DT) as their baseline classifiers.

The second stage involves rebalancing imbalanced medical data using four strategies: the SMOTE (Chawla et al., 2002), cost-sensitive learning (Elkan, 2013), and random oversampling and undersampling techniques (Japkowicz, 2000). Similarly, when training with the base classifiers selected from the previous stage, the framework suggests determining which of the four strategies yields the most substantial performance improvements across various parameter configurations for each rebalancing strategy. The available hyperparameter tuning range must be developed using criteria for each rebalancing strategy's parameter/threshold.

As Zhao et al. (2018) suggested, datasets with imbalanced distributions need to be rebalanced (using each strategy) and validated using the LOOCV in stage three. A conventional cross-validation technique in which one sample is excluded (leave-out) for validation and training is performed on the other samples supplied to the model; this

procedure is repeated on all samples. LOOCV is widely favored by many researchers for extensive validation processes, where the number of cross-validations is determined by the number of instances in a dataset (Cheng et al., 2017). In the medical domain, it has been implemented in many model validations, such as the biomedical phenotype predictive model (deAndrés-Galiana et al., 2016), Alzheimer's classification model (Cuingnet et al., 2011), breast cancer model (Liang et al., 2018), and kidney stone predictive model (Shabaniyan et al., 2019).

Table 1 summarizes the adapted framework process by Zhao et al. (2018) in stages. The framework provides detailed insight into the framework we adapted for our experiments.

Table 1
Stages of the adapted (Zhao et al., 2018) rebalancing framework process

Stage	Process	Description	Selections
1 Classifier • R.SVM baseline	Selecting base linear or non-linear	Candidate classifiers can be either Select the classifier that performs best as	• LR • L.SVM • DT
2 Rebalancing strategies • Strategy parameter/ threshold • Sensitive learning	Incoporating strategies according to results of each parameter tune	Select best-performed rebalancing Oversampling Tune according to each rebalancing	• SMOTE • Random • Random undersampling cost
3 Validate model		Estimate finalized model performance	• LOOCV

MATERIALS AND METHODS

Ensemble-based Classifier

A unified classifier overcomes the limitations of each counterpart in terms of accuracy and performance (Utami et al., 2014). As mentioned earlier, we employed the hard-level technique using majority voting for classifier combinations in this study. This approach combines the highest predicted class output from each classifier. For instance, if six out of eleven classifiers vote for the same class output, the class with the highest number of votes is considered the final result. The formulation for our hard-level majority voting, *Em*, is calculated using Equation 1:

$$Em = \sum_{i=1}^{M} d_{i,k} = \max_{j=1}^{h} \sum_{i=1}^{M} d_{i,j} \quad [1]$$

where M is the total number of classifiers and h is the total number of classes. However, the class that received the same maximal vote (tie) can be resolved using the weighted

majority voting (Kuncheva, 2014) to choose the class with higher weighted votes. The weighted majority voting is calculated using Equation 2:

$$\sum_{i=1}^{M} b_i d_{i,k} = \max_{j=1}^{h} \sum_{i=1}^{M} b_i d_{i,j} \quad [2]$$

where b_i is the weighting coefficient for classifier D_i.

Self-paced Ensemble

Classifiers tend to prioritize a class with more samples when learning from highly skewed data, leading to biased predictions. Consequently, the ability of classifiers to distinguish between minority and majority classes is highly dependent on the data distribution they learn. Conventional rebalancing techniques (e.g., random oversampling, undersampling, SMOTE) offer common approaches for imbalanced learning. Going further, Liu et al. (2020) introduce a novel imbalance learning method, the Self-paced Ensemble (SPE). SPE incorporates the concept of 'hardness' in classification, describing the difficulty of categorizing a sample for a given classifier. Derived from this difficulty, SPE systematically chooses the most informative data samples that align with their distribution. Equation 3 calculates the hardness:

$$H(x, y, F) = \frac{1}{n}\sum_{i=1}^{n} |f_i(x_i) - y_i| \quad [3]$$

where H is the hardness function, F can be any chosen classifier and dataset as (x,y). $F(x_i)$ indicates the classifier's probability. Liu et al. (2020) state that SPE can be adapted to any classifier. As mentioned previously, to align with the use case of this study, we adapted our proposed ensemble-based classifier, *Em*, as the SPE base estimator to enhance its effectiveness. The new adapted hardness function with Em is defined in Equation 4:

$$SPE(Em) = H(x, y, Em) = \frac{1}{n}\sum_{i=1}^{n} |f_i(x_i) - y_i| \quad [4]$$

SPE enhances the significance of boundary samples by incorporating an undersampling technique to reduce the presence of noisy and insignificant data samples. It is achieved by dividing most samples into k bins based on their hardness rating, where k is the hyperparameter. Each bin is then undersampled to create a balanced dataset, ensuring that every bin has similar hardness. The formulation of SPE with our adapted Em, SPE(*Em*), is shown in Equation 5:

$$B_l = \left\{(x, y) \middle| \frac{(l-1)}{k} SPE(Em) \leq \frac{l}{k}\right\} w.l.o.g.H \in [0,1] \quad [5]$$

where B_l is used to denote the l-th bin.

Performance Evaluation Metrics

The most commonly used metrics for classifier model performance are accuracy, precision, and recall. Accuracy represents the overall proportion of correctly predicted instances across all classes, calculated using Equation 6:

$$Accuracy = (TP + TN)/(TP + FP + TN) \quad [6]$$

where TP is the true positive, TN is the true negative, and FP is the false negative.

Meanwhile, precision focuses on the true positive rate within the positive predictions, while recall measures the ability of classifiers to correctly identify the actual positive class (Grandini et al., 2020). Equation 7 calculates the precision, while Equation 8 calculates the recall:

$$Precision = (TP)/(TP + FP) \quad [7]$$

$$Recall = (TP)/(TP + FN) \quad [8]$$

where FN is the false negative.

However, accuracy alone may not offer a comprehensive view of a classifier's performance in class imbalance due to the bias inherent in the class distribution between the minority and majority classes. High precision may come at the cost of low recall and vice versa when it comes to precision and recall. Maintaining an appropriate balance between these metrics becomes crucial for effectively handling imbalanced data. Therefore, depending solely on these metrics in an imbalanced class scenario can be misleading (Akosa, 2017).

In this study, we have incorporated F-score and G-mean as additional metrics to obtain a more accurate and comprehensive assessment in such scenarios. The F-score, in particular, offers a balanced evaluation that takes into account both precision and recall, as it offers a well-rounded assessment of a classifier's performance. It considers FPs and FNs to determine the harmonic mean of precision and recall (Phoungphol et al., 2012). Equation 9 calculates the F-score:

$$F - score = \frac{(Recall \times Precision)}{(Recall + Precision)} \quad [9]$$

The G-mean metric offers valuable insights into a classifier's capability to classify minority class instances, a crucial metric for class imbalance. Additionally, it considers both TPs and TNs, ensuring a comprehensive evaluation of a classifier's performance with equal weight given to both classes. Consequently, it prevents excessive bias toward the majority class, fostering a more balanced approach. (Błaszczyk & Jedrzejowicz, 2021). Equation 10 calculates the G-mean:

$$G - mean = \sqrt{Recall \times Precision} \quad [10]$$

Rebalancing Strategies

As per the recommendation by the adapted framework, this study incorporates four rebalancing strategies: random oversampling, random undersampling, SMOTE, and CSL. Random oversampling involves duplicating instances from the minority class to balance class distribution (Barua et al., 2014). This strategy is effective when the dataset has a small number of minority class instances, but it may lead to overfitting if not applied cautiously. Conversely, random undersampling randomly reduces the number of majority class instances to match the minority class, making it suitable for datasets with a large majority class and when computational efficiency is a concern. However, it may lead to the loss of valuable information from the majority class (Barua et al., 2014).

SMOTE generates synthetic instances in the minority class by interpolating between existing instances, thereby enhancing the representation of the minority class and achieving a balanced class distribution while removing bias. It is commonly employed in various imbalance learning scenarios and proves particularly useful when limited data is available for the minority class (Kotsiantis et al., 2006). CSL, on the other hand, assigns different misclassification costs to different classes, emphasizing the importance of the minority class. This approach is beneficial in cases of severe class imbalance, where misclassifying the minority class carries higher consequences (Krawczyk, 2016). CSL aims to reduce the misclassification of the minority class by making it more costly for the classifier.

The rationale for choosing which rebalancing strategy to use depends on the specific characteristics of the dataset, such as the class distribution, dataset size, and the consequences of misclassification. While one strategy may yield enhanced performance for a certain dataset, it might not prove as effective for another. The choice of strategy can also be influenced by the type of classifier used, as different classifiers may interact differently with rebalancing techniques (Cipriano et al., 2021), leading to varied performance outcomes. Therefore, this study independently implemented each strategy to identify the most effective strategy for the specific dataset.

Experimental Setup

In our first experiment, we assessed the overall performance of the rebalancing framework outlined by Zhao et al. (2018) on each dataset. It entailed using the initial baseline classifiers recommended within the framework: LR, L.SVM, R.SVM, and DT. We directly applied these classifiers, compared their results, and identified the top-performing classifier as our baseline. This experimental approach is denoted as ‘Experiment 1’.

In our second experiment, rather than comparing individual candidate classifiers, we amalgamated them into a unified classifier using hard-level majority voting. Following this, we applied the framework utilizing the ensemble baseline classifier to assess its performance. This experimental approach is designated as ‘Experiment 2’.

Finally, in the third experiment, we employed our proposed SPE with an ensemble-based classifier as the base estimator, denoted as SPE(*Em*). *Em* represents a combination of DT C4.5, R.SVM, and XGB. This experimental approach is designated as 'Experiment 3'.

The DT c4.5 algorithm is a conventional yet powerful classification method frequently used to solve medical diagnosis problems (Breiman, 2001). Radial is a well-known kernel function that is utilized in a variety of kernelized learning techniques. It is part of a kernel function embedded in the standard support vector machine (SVM). Hence, the name R.SVM. Kernels is the application-specific measure of similarity between data instances used by SVM. R.SVM proved to show significant classification performance in the medical domain for predicting diseases (Harimoorthy & Thangavelu, 2021). Meanwhile, XGB is a more regularized, expanded version of a gradient boosting method that provides a robust boosted tree model with high accuracy and is known for its ability to classify imbalanced datasets (Cahyana et al., 2019; Ma et al., 2022).

Finally, we compared the performance results obtained from each experiment to ascertain which experiment yielded robust overall performance across all datasets. Five evaluation performance metrics were used to measure the performance in each experiment: accuracy, precision, recall, F-score, and G-mean to measure the performance in each experiment. Additionally, we incorporated the receiver operating characteristic curve (ROC) and root mean square error (RMSE) as part of our evaluation metrics. These metrics are commonly used to evaluate the performance of a model in a binary classification. The overall workflow is shown in Figure 1. Our experiment setup codes are available on GitHub (https://tinyurl.com/vxphztfe).

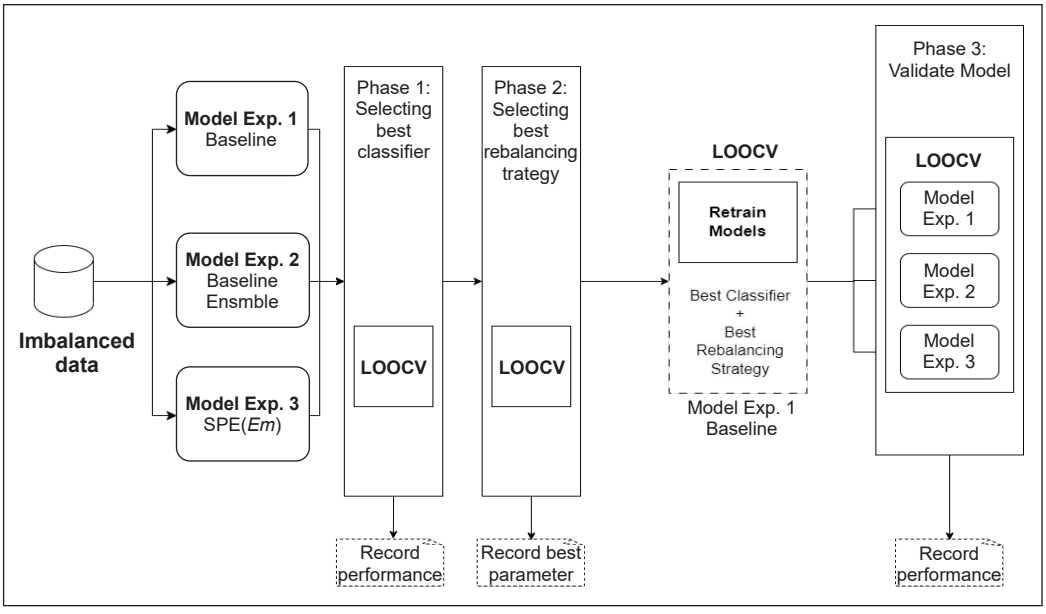


Figure 1. Workflow of experimental approaches

Statistical Test

In this study, we employed the paired t-test as the preferred statistical analysis method to assess the significance of the results. The paired t-test is a well-known statistical test that allows us to compare the means of two related groups while taking into account the dependency between them (Newcombe, 1992). Additionally, it assesses whether there is a significant difference in the means of paired observations while taking into account that the observations are dependent. It does so by calculating a t-statistic that measures the standardized difference between the means of the paired observations.

In the case of the paired t-test, it is calculated based on the differences between paired observations. Meanwhile, the p-value associated with the t-statistic indicates the likelihood of observing such a difference by chance. A lower p-value indicates a higher degree of statistical significance. If the p-value is below a predetermined significance level (<0.05), we can infer a statistically significant difference in the performance outcomes between the experiments. By employing the paired t-test, we aimed to rigorously assess the statistical significance of the improvements observed in Experiment 3, thus providing robust evidence of the effectiveness of our proposed ensemble-based classifier.

Datasets

Our experiments utilized seven imbalanced medical datasets from the UC Irvine machine learning repository (UCI), Kaggle, and Knowledge Extraction based on Evolutionary Learning (KEEL). These datasets are identified as Heart disease (Cleveland0vs4), eColi4, Yeast3, SPECT, SPECTF, Parkinson, and Cirrhosis. Each dataset exhibits a distinct level of imbalanced class distribution. The datasets are structured in a tabular and binary format. We calculated the class ratios for each dataset, representing the level of imbalance as the IR. A higher IR indicates a more imbalanced distribution (Zhu, Guo, & Xue, 2020). Equation 11 is used to calculate the IR for a binary class problem:

$$IR = \frac{N_{max}}{N_{min}} \quad [11]$$

where N_{maj} is the number of majority instances, and N_{min} is the number of minority instances.

Note that in this study, our focus is solely on binary classification. We selected datasets initially formatted for binary classification to maintain the binary setting. Table 2 summarizes the structure of these datasets. It shows that the Cirrhosis dataset has the highest level of imbalance, with IR = 18.90. It is followed by the eColi4 and Cleveland0vs4 datasets, with IR = 15.80 and 12.62, respectively. The others have IR below 8.10, and Parkinson has a minimum IR=3.06.

Table 2
Summary of imbalanced medical datasets

Dataset	No. of Records	No. of Features	Class Distribution			Imbalance Ratio (N_{maj}/N_{min})	Source
			Class	Samples	Percentage (%)		
Cleveland0vs4	177	13	0	164	92.65%	12.62	KEEL[1]
			1	13	7.35%		
eColi4	336	7	0	316	92.65%	15.80	KEEL[1]
			1	20	5.95%		
Yeast3	1484	8	0	1321	89.02%	8.10	Kaggle[2]
			1	163	10.98%		
SPECT	267	22	0	55	20.60%	3.85	UCI[3]
			1	212	79.40%		
SPECTF	267	44	0	55	20.59%	3.85	UCI[3]
			1	212	79.41%		
Parkinson	195	22	0	48	24.61%	3.06	UCI[3]
			1	147	75.39%		
Cirrhosis	418	13	0	21	5.02%	18.90	Kaggle[2]
			1	397	94.98%		

[1]http://www.keel.es; [2]https://www.kaggle.com/; [3]http://archive.ics.uci.edu/ml

RESULTS

To facilitate readers' understanding of the experimental part, we run each experiment according to the process explained in the previous discussion. We then discuss and compare the results with and without the rebalancing strategy applied.

Experimental Results

We executed the original framework from Zhao et al. (2018) across all the imbalanced datasets. The main focus was to assess the performance of each candidate classifier recommended by the framework and identify the classifier that demonstrated the highest performance on each dataset. In our experiments, we performed LOOCV for stages 1,2 and 3. We then record the performance of each stage. The results for stages 1 and 3 are shown in Table 3, which shows the average LOOCV results of all experimental approaches with and without the rebalancing strategy applied. Meanwhile, the results for stage 2 are shown in Table 4.

The analysis based on the experimental results is as follows:

1. According to Table 3, the datasets highlighted in bold demonstrated the best overall performance on each dataset in terms of accuracy, precision, recall, F-score, and G-mean. The left side of the table (no rebalancing) shows that all models have relatively acceptable accuracy but low values for the other metrics. The imbalanced nature of data distribution contributes to this degradation, especially in Cirrhosis,

Table 3
Comparison of average LOOCV classification performance with/without rebalancing strategy

Dataset	Exp.	Classifier	No Rebalancing						With Rebalancing					
			Acc	Prec	Rec	F-score	G-mean	TP/FN/FP/TN	Acc	Prec	Rec	F-score	G-mean	TP/FN/FP/TN
Clevelan0vs4	Exp 1	L.SVM	93.22%	53.85%	53.85%	53.85%	72.03%	7/6/6/158	85.31%	32.43%	92.31%	48%	88.45%	12/1/25/139
	Exp 2	Base Em.	93.79%	62.5%	38.46%	47.62%	61.45%	5/8/3/161	95.48%	72.73%	61.54%	66.67%	77.73%	8/5/3/161
	Exp 3	**SPE(*Em*)**	**94.35%**	**57.89%**	**84.62%**	**68.75%**	**89.71%**	**11/2/8/156**	**93.79%**	**54.55%**	**92.31%**	**68.77%**	**93.10%**	**12/1/10/154**
eColi4	Exp 1	DT	96.13%	65.22%	75%	67.77%	85.50%	15/5/8/308	97.62%	87.50%	70%	77.78%	83.40%	14/6/2/314
	Exp 2	Base Em.	98.81%	110%	80%	88.89%	89.44%	16/4/0/316	98.51%	100%	75%	85.71%	86.60%	15/5/0/316
	Exp 3	**SPE(*Em*)**	**98.21%**	**85%**	**85%**	**85%**	**91.76%**	**17/3/3/313**	**98.21%**	**81.82%**	**90%**	**85.71%**	**94.27%**	**18/2/4/312**
Yeast3	Exp 1	L.SVM	94.61%	78.23%	70.55%	74.19%	82.97%	115/48/32/1289	93.87%	69.15%	79.75%	74.07%	87.32%	130/33/58/1263
	Exp 2	Base Em.	94.61%	82.17%	65.03%	72.60%	79.94%	106/57/23/1298	94.95%	74.72%	81.60%	78.01%	88.78%	133/30/45/1276
	Exp 3	**SPE(*Em*)**	**95.15%**	**75.71%**	**82.21%**	**78.82%**	**89.18%**	**134/29/43/1278**	**94.61%**	**72.93%**	**80.98%**	**76.74%**	**88.31%**	**132/31/49/1272**
SPECT	Exp 1	L.SVM	82.77%	88.43%	90.09%	89.25%	70.10%	191/21/25/30	76.40%	91.16%	77.83%	83.97%	74.29%	165/47/16/39
	Exp 2	Base Em.	81.27%	88.21%	88.21%	88.21%	69.36%	187/25/25/30	74.53%	92.35%	74.06%	82.20%	75.20%	157/55/13/42
	Exp 3	**SPE(*Em*)**	**76.78%**	**94.12%**	**75.47%**	**83.77%**	**78.58%**	**160/52/10/45**	**79.40%**	**91.53%**	**81.60%**	**86.28%**	**76.07%**	**173/39/16/39**
SPECTF	Exp 1	L.SVM	79.78%	86.57%	88.21%	87.38%	64.57%	187/25/29/26	73.78%	88.17%	77.36%	82.41%	68.13%	164/48/22/33
	Exp 2	Base Em.	80.90%	87.44%	88.68%	88.06%	67.19%	188/24/27/28	69.29%	92.76%	66.51%	77.47%	72.94%	141/71/11/44
	Exp 3	**SPE(*Em*)**	**73.03%**	**94.87%**	**69.81%**	**80.43%**	**77.24%**	**148/64/8/47**	**77.15%**	**90.37%**	**79.72%**	**84.71%**	**73.23%**	**169/43/18/37**
Parkinson	Exp 1	L.SVM	87.69%	87.73%	97.28%	92.26%	75.33%	143/4/20/28	80.51%	90.37%	82.99%	86.52%	77.79%	122/25/13/35
	Exp 2	Base Em.	87.69%	88.20%	96.60%	92.21%	76.39%	142/5/19/29	78.46%	93.39%	76.87%	84.33%	80.04%	113/34/8/40
	Exp 3	**SPE(*Em*)**	**92.31%**	**95.21%**	**94.56%**	**94.88%**	**89.87%**	**139/8/7/41**	**93.85%**	**94.70%**	**97.28%**	**95.97%**	**90.04%**	**143/4/8/40**
Cirrhosis0vs4	Exp 1	L.SVM	94.98%	94.98%	100%	97.42%	0%	397/0/21/0	76.56%	96.00%	78.59%	86.43%	54.72%	312/85/13/8
	Exp 2	Base Em.	94.74%	94.96%	99.75%	97.30%	0%	396/1/21/0	78.95%	97.54%	79.85%	87.81%	70.31%	917/80/8/13
	Exp 3	**SPE(*Em*)**	**62.44%**	**98%**	**61.71%**	**75.73%**	**68.57%**	**245/152/5/16**	**84.21%**	**97.42%**	**85.64%**	**91.15%**	**69.96%**	**340/57/9/12**

Acc= Accuracy, Prec = Precision, Rec = Recall

eColi4, and Cleveland0vs4. The ensemble baseline in Experiment 2 performed poorly due to the individual weak learners underachieving the final output. However, SPE(*Em*) in Experiment 3 achieved considerably adequate balance results in all the performance measures across most of the dataset. For instance, it has a higher G-mean in Cirrhosis with 68.57%, 89.71% for Cleveland0vs4, and 91.76% for eColi4.

2. Table 3, on the right side of the table (with the best rebalancing strategy), shows that all experimental models substantially improved overall performance across all the metrics. Rebalancing data contributes to better classifier performance.
3. Comparing the experimental results in Table 3, SPE(*Em*) outperforms the other experiment models with significant improvements, followed by Experiment 2 and Experiment 1. For instance, SPE(*Em*) achieved an increase of F-score from 75.73% to 91.15% in the Cirrhosis dataset after rebalancing with the best strategy. Experiment 2 slightly outperforms Experiment 1 in terms of f-score and G-mean, especially in Cleveland0vs4, eColi4, Yeast3, and Cirrhosis. Notice that the results for Experiment 3 have adequate performance even without rebalancing and achieved slightly better performance after rebalancing.

Table 4 compares the best-selected base classifier and rebalancing strategy with the best performance. For SMOTE, the number of oversampled minority instances, α, and undersampled majority instances, γ, are controlled by these two parameters, respectively. SMOTE handles imbalanced datasets by oversampling the minority class. Even if the examples provide no new information to the model, SMOTE will duplicate the instances from the minority class and construct new instances by synthesizing the existing examples (Chawla et al., 2002). We applied SMOTE during the LOOCV to resample each training fold and validate on the test fold; the same approach also applies to the random sampling (under and oversampling) method.

Note that for each rebalancing strategy, we only applied the parameter settings that showed the highest improvement in overall results. Table 4 shows that Cost-sensitive

Table 4
A comparison of the selected base classifier and rebalancing strategy gives the best performance with LOOCV

Experiment	Dataset	Best Base Classifier	Best Rebalancing Strategy	
			Best Strategy	Best Parameter Setting
Exp 1	Cleveland0vs4	L.SVM	SMOTE	$\alpha = 0.68, \gamma = 0.85$
	eColi4	DT	Oversampling	ratio = 0.85
	Yeast3	L.SVM	Oversampling	ratio = 0.35
	SPECT	L.SVM	SMOTE	$\alpha = 0.5, \gamma = 0.5$
	SPECTF	L.SVM	Oversampling	ratio = 0.68
	Parkinson	L.SVM	SMOTE	$\alpha = 0.75, \gamma = 0.35$
	Cirrhosis0vs4	L.SVM	Undersampling	Ratio = 0.5

Table 4 *(continue)*

Experiment	Dataset	Best Base Classifier	Best Rebalancing Strategy	
			Best Strategy	Best Parameter Setting
Exp 2	Cleveland0vs4	Baseline Ensemble (LR+L.SVM+ R.SVM+DT)	Oversampling	ratio = 0.85
	eColi4		CSL	Threshold = 0.015
	Yeast3		Oversampling	ratio = 0.5
	SPECT		CSL	Threshold = 0.25
	SPECTF		SMOTE	$\alpha = 0.65, \gamma = 0.65$
	Parkinson		SMOTE	$\alpha = 0.5, \gamma = 0.5$
	Cirrhosis0vs4		CSL	Threshold = 15
Exp 3	Cleveland0vs4	SPE(*Em*)	Undersampling	ratio = 0.15
	eColi4		CSL	Threshold = 0.5
	Yeast3		CSL	Threshold = 0.015
	SPECT		Undersampling	ratio = 0.28
	SPECTF		SMOTE	$\alpha = 0.65, \gamma = 0.65$
	Parkinson		Oversampling	ratio = 0.75
	Cirrhosis0vs4`		Oversampling	Ratio = 0.1

learning (CSL), oversampling and SMOTE performed best with most of the datasets in all experiments. Most datasets in Experiment 1 were best rebalanced with oversampling, while Experiments 2 and 3 favored the other sampling methods.

Overall Results

Per the framework's recommendation, we performed the LOOCV with varying repetitions according to the number of instances on each dataset (Zhao et al., 2018). We obtained the results and recorded them as the average validation performance across all datasets by an experimental approach. Table 5 presents a comparative analysis of overall performance metrics, as average LOOCV across all datasets for each experimental approach.

According to Table 5, Experiment 1 with baseline classifier (LR, L.SVM, R.SVM and DT) performs at 84.10%, 79.34%, 80.71%, 77.53%, and 76.70% in terms of accuracy, precision, recall, F-score, and G-mean, respectively. In Experiment 2, there was a noticeable increase in F-score and G-mean by 81.36% and 79.70%; however, recall declined to 75.32%. Meanwhile, the average performance obtained in Experiment 3 was 89.06%, 85.23%, 86.39%, 85.38%, and 83.54% in terms of accuracy, precision, recall, F-score, and G-mean, respectively. Overall, the proposed method used in Experiment 3 has improved the overall performance compared with the baseline conditions (Experiment 1) by 4.96%, 5.89%, 5.68%, 7.85%, and 6.84% in terms of accuracy, precision, recall, F-score, and G-mean respectively. The results also show that the combined baseline classifier in Experiment 2 has increased performance and is slightly better than Experiment 1, especially in the F-score of 81.36%.

Table 5
Average LOOCV performance comparison on all datasets for each experimental approach

xp.	Classifier	Performance Achieved				
		Acc	**Prec**	**Rec**	**F-score**	**G-mean**
Exp 1	Baseline: LR, L.SVM,R.SVM,DT	84.10%	79.34%	80.71%	77.53%	76.70%
Exp 2	Baseline En	85.04%	89.14%	75.32%	81.36%	79.70%
Exp 3	**SPE(*Em*)**	**89.06%**	**85.23%**	**86.39%**	**85.38%**	**83.54%**
	SPE(*Em*) Increase from Exp 1	**+4.96%**	**+5.89%**	**+5.68%**	**+7.85%**	**+6.84%**

Acc= Accuracy, Prec = Precision, Rec = Recall

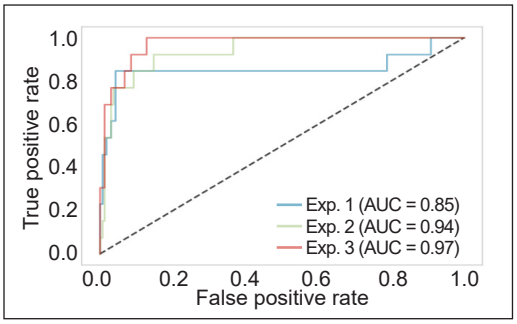
The ROC, which plots the true positive rate (TPR) against the false positive rate (FPR) for each experiment, is depicted in Figure 2. These are used to assess the robustness of the three experimental approaches. Applied by many researchers, ROC curves are a useful way to evaluate imbalanced data (Turlapati & Prusty, 2020; Phoungphol et al., 2012; Yao & Chen, 2019). We performed the ROC analysis with the LOOCV. The area under ROC (AUROC) curves for each approach are shown in blue, green, and red for Experiments 1, 2 and 3, respectively, in Figure 2. The performance of a 'random guessing classifier' for the class of observations is depicted by the grey dashed line in each figure (no-discrimination line). A successful classification technique should provide points close to or in the top part of the graph (0,1) (Saito & Rehmsmeier, 2015; Mandrekar, 2010).

All plots of TPR versus FPR lie above the grey line, indicating that all three approaches are able to handle the binary class classification problem. However, the ROC for SPE(*Em*) in Experiment 3 is closer to coordinate (0,1) on Cleveland0vs4, eColi4, SPECT, SPECTF, Parkinson, and Cirrhosis with AUC 0.97, 0.99, 0.83, 0.84, 0.96, and 0.73, respectively. However, all experiments achieved a similar AUC of 0.97 on the Yeast3 dataset, with Experiment 3 having a slightly closer curve, followed by the second-best model in Experiments 2 and 1. Decisively, Experiment 3 demonstrates higher ROC results than the other two experiments across most of the datasets, indicating the model could significantly distinguish between the positive and negative classes for better classification.

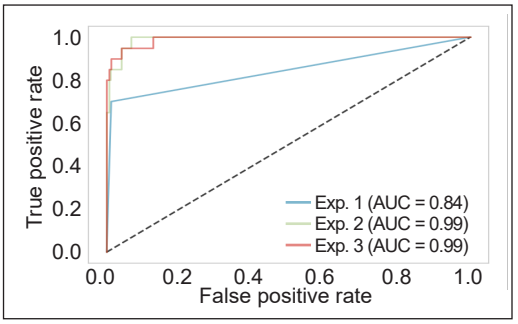
We also record the RMSE on each fold of LOOCV with respect to each dataset to evaluate the error rate. The square root of MSE is referred to as RMSE. The error rate is a percentage measure of the difference between the actual and estimated values. The lower the RMSE (>=0), the lower the error rate. The RMSE is reported in Table 6.

Table 6
RMSE of all experimental models on each dataset

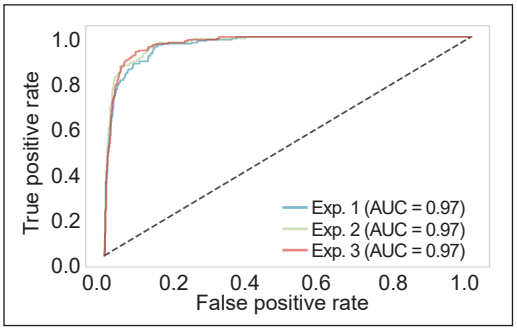
Dataset	Exp 1	Exp 2	Exp 3
	RMSE	*RMSE*	*RMSE*
Cleveland0vs4	0.36	0.21	0.24
eColi4	0.17	0.15	0.14
Yeast3	0.22	0.2	0.21
SPECT	0.57	0.4	0.4
SPECTF	0.4	0.42	0.43
Parkinson	0.35	0.32	0.25
Cirrhosis	0.23	0.21	0.19
Average	0.3286	0.2729	**0.2657**



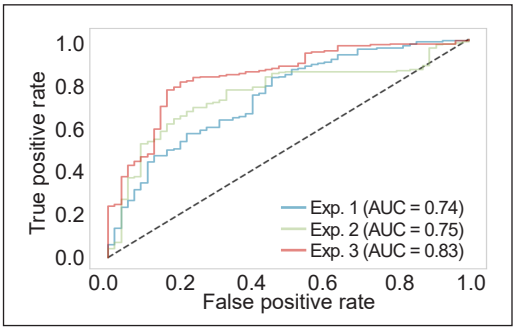
(a)



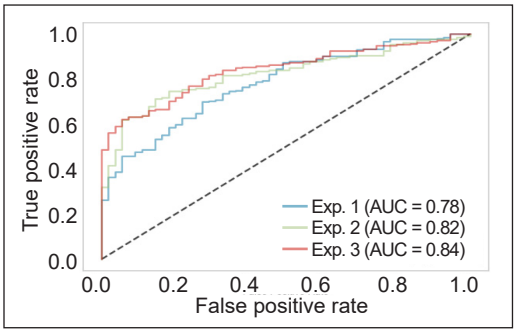
(a)



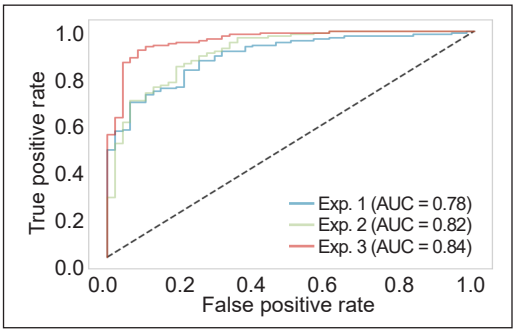
(c)



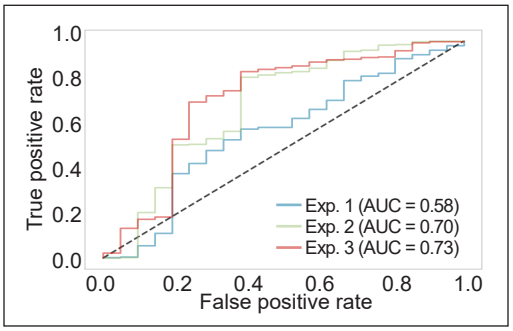
(d)



(e)



(f)



(g)

Figure 2. ROC curves for each dataset with different experimental approaches: (a) Cleveland0vs4; (b) eColi4; (c) Yeast3; (d) SPECT; (e) SPECT; (f) Parkinson; and (g) Cirrhosis

As shown in Table 6, the SPE(*Em*) in Experiment 3 performed best with RMSE values of 0.14, 0.25, and 0.19 for eColi4, Parkinson's, and Cirrhosis, respectively. Averaged at 0.2657 overall RMSE. While Experiment 2 has an RMSE average of 0.2729, it is slightly closer to Experiment 3. Experiment 1 was performed at an RMSE value of 0.3286, distinctly higher than the other experiments. Comparing the experimental models, SPE(*Em*) seems to have a slight edge over the Experiment 2 model.

Table 7
Paired t-test statistical results of Experiment 3 with the other two experimental approaches

	Paired Differences					t	df	p-value
	Mean	Std. Deviation	Std. Error mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Exp 3 – Exp 1	6.24	1.0026	0.4484	4.990	7.4890	12.4550	4	.000239
Exp 3 – Exp 2	3.81	4.7412	2.1203	-2.079	9.6950	1.6063	4	.1835

Table 7 shows the statistical test results using the paired t-test between Experiment 3 and the other two experimental approaches. The comparison between Experiment 3 and Experiment 1 revealed a significant difference in performance. Experiment 3 displayed a substantial improvement with a mean difference of 6.24, a low standard error (0.4484), and a narrow confidence interval (4.990 to 7.4890). Additionally, the high t-statistic (12.4550) and the extremely low p-value (0.000239) emphasized the statistical significance of Experiment 3's superior performance over Experiment 1. Meanwhile, the comparison between Experiment 3 and Experiment 2 was not statistically significant due to its p-value of 0.1835, which is more than the significant level of 0.05. However, it is important to note that both Experiment 3 and Experiment 2 utilized the same ensemble method. It highlights the effectiveness of the ensemble approach, as both Experiment 3 and Experiment 2 consistently outperformed Experiment 1, the baseline method. Although statistical significance may not be established in every case, the shared use of the ensemble method highlights its effectiveness in enhancing overall performance.

DISCUSSION

This study presents our investigation into the performance of ensemble-based classifiers within the Zhao et al. (2018) framework, employing seven imbalanced datasets. Our experimental results clearly indicate that rebalancing methods enhance the overall predictive learning of classifiers (Table 3). To evaluate the performance of each experimental model, we LOOCV for Stages 1 to 3 and recorded the results. Significantly, the performance of each model improved, with SPE(*Em*) in Experiment 3 demonstrating the best overall performance, followed by Experiment 2 and *1*. Our primary metrics for imbalance learning

are AUROC, F-score, and G-mean. Specifically, the F-score is apt for discriminating between the minority and majority classes. AUROC summarizes a model's capacity to discriminate between classes, and G-mean measures the minority class performance. In terms of overall performance (Table 5 and Figure 2), our proposed SPE(*Em*) yielded significant results for all three metrics, followed by the baseline ensemble in Experiments 2 and 1.

SPE(*Em*) also improved the overall performance compared to the baseline by 4.96%, 5.89%, 5.68%, 7.85% and 6.84% in accuracy, precision, recall, F-score, and G-mean, respectively. The improved performance observed in Experiments 2 and 3 is attributed to utilizing an ensemble of classifiers, particularly stronger classifiers capable of mitigating class imbalances. It aligns with findings from prior studies (Jiang et al., 2020; Valentini & Dietterich, 2004) that demonstrate how incorporating ensemble methods leads to a unified improvement in overall performance. In Experiment 3, boosting the performance of the ensemble classifier (R.SVM, DT, and XGB) with SPE showed increased results. It is also relevant to point out that the ensemble-based classifiers can achieve consistent and stable performance with increased results compared to the baseline (Experiment 1). Therefore, classifying imbalanced data proves to have a significant impact on the objective of this study.

Experiment 1 served as the essential baseline for comparison with the other two experiments; hence, we refer to it as the benchmark experiment. By comparing the results of this benchmark experiment, we have demonstrated that the proposed ensemble-based classifier in Experiment 3 achieved superior outcomes. The findings from both Experiments 2 and 3 offer compelling evidence of the effectiveness of ensemble-based classifiers in enhancing the existing framework (Zhao et al., 2018). Consequently, the results from Experiment 3 will serve as the cornerstone for our future endeavors in developing a rebalancing framework integrated with ensemble-based classifiers soon.

The results of these experiments are in line with Zhao et al. (2018), which further supports its applicability on various applications not just limited to medical incident reports but also various medical data with similar class imbalanced properties. Additionally, it is also worth mentioning that these results are consistent with previous studies implementing ensemble-based classifiers to address class imbalances in medical data (Zhu et al., 2015; Sandhan & Choi, 2014). Notably, prior similar works (Krishnan & Sangar, 2021; Song et al., 2022; Bi & Ma, 2021; Tang et al., 2021) incorporating ensemble methods in their rebalancing frameworks have shown significant results, further supporting the effectiveness of ensembles in handling class imbalances. That said, this study is not a replacement for the original framework by Zhao et al. (2018); instead, it provides ample insight and opportunities for researchers to explore more ensemble disciplines in addressing class imbalanced problems in different domains. Future studies may still implement Zhao et al.'s

(2018) for class imbalance; however, with the results of this experiment, our future works will involve a new multi-class rebalancing framework incorporating an ensemble method.

LOOCV was used to evaluate the finalized model, as recommended by Zhao et al.'s (2018) framework. The downside of using LOOCV is that it requires a high time complexity, depending on the number of replications applied. Nonetheless, due to the small dataset used by the authors (Zhao et al., 2018), this issue was inconsequential to their research and was neglected. However, this is not the case in our experiment since our datasets have varying sizes (especially Yeast3). Despite the time complexity concern, we opted to use LOOCV to ensure a comparative analysis of Zhao et al.'s (2018) framework with minimal bias. In future works, we may explore using k-fold cross-validation as a more convenient method for estimating model performances.

Furthermore, it is worth noting that this study is limited to imbalanced binary classification problems, as was true in the previous authors' results (adapted framework). It ensures a fair comparison while maintaining fidelity to the framework and avoiding potential bias.

In this study, we investigated the imbalanced nature of the medical dataset with a state-of-the-art rebalancing framework combined with our proposed ensemble approach (Zhao et al., 2018). However, we also observed that class imbalance is not the sole issue in the medical domain. Another complicating factor is the limited availability of data. Due to strict privacy regulations and data-sharing constraints, medical data has become scarce. Consequently, many machine learning researchers resort to publicly available medical datasets. For this reason, we could not obtain more medical datasets with high dimensionality; hence, the small sample size dataset (Cleveland0vs4, Parkinson, SPECT, and SPECTF) was used in our experiment. Since our focus is on imbalanced learning, these seven publicly available datasets proved sufficient for this study. However, we plan to explore state-of-the-art synthetic data generation methods in future works, such as Generative Adversarial Networks (GAN) and Variational Autoencoders (VAE). These techniques offer the advantage of creating synthetic data that closely resembles real-world data (Abedi et al., 2022; Elbattah et al., 2021).

CONCLUSION

This paper comprehensively analyses an ensemble-based classifier within a rebalancing framework for imbalanced medical data. Our experimental results demonstrate significant performance improvements, particularly incorporating SPE(*Em*) in Experiment 3. The effectiveness of ensemble-based classifiers in addressing class imbalances is highlighted, with consistent performance enhancements observed across experimental approaches. Additionally, imbalanced data are prevalent in the medical domain, encompassing binary and multi-class classification scenarios. Although this study is limited to binary

classification, it becomes evident that the issue also exists in the context of multi-class classification. Therefore, we intend to develop a multi-class rebalancing framework incorporating an ensemble-based classifier to address the challenges of multi-class imbalanced datasets in the medical domain.

ACKNOWLEDGEMENT

The authors thank the Ministry of Higher Education Malaysia and Universiti Teknologi MARA, Malaysia, for financially supporting this project under the Fundamental Research Grant Scheme (FRGS) Grant No. FRGS/1/2023/ICT01/UITM/02/4. The authors would also like to thank the College of Computing, Informatics and Mathematics and Faculty of Medicine, Universiti Teknologi MARA, Selangor, Malaysia, for all the support.

REFERENCES

Abedi, M., Hempel, L., Sadeghi, S., & Kirsten, T. (2022). GAN-based approaches for generating structured data in the medical domain. *Applied Sciences*, *12*(14), Article 7075. https://doi.org/10.3390/app12147075

Abraham, A., & Elrahman, S. M. A. (2013). A review of class imbalance problem. *Journal of Network and Innovative Computing*, *1*, 332–340.

Belarouci, S., & Chikh, M. A. (2017). Medical imbalanced data classification. *Advances in Science, Technology and Engineering Systems*, *2*(3), 116–124. https://doi.org/10.25046/aj020316

Bi, W., & Ma, R. (2021). Unbalanced data set processing method for colorectal cancer prediction in TCM diagnosis. In *2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM)* (pp. 1-6). IEEE Publishing. https://doi.org/10.1109/HEALTHCOM49281.2021.9615914

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cahyana, N., Khomsah, S., & Aribowo, A. S. (2019). Improving imbalanced dataset classification using oversampling and gradient boosting. In *2019 5th International Conference on Science in Information Technology (ICSITech)* (pp. 217-222). IEEE Publishing. https://doi.org/10.1109/ICSITech46713.2019.8987499

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*(2), 321–357. https://doi.org/10.1613/jair.953

Cheng, H., Garrick, D. J., & Fernando, R. L. (2017). Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction. *Journal of Animal Science and Biotechnology, 8*(1), 1–5. https://doi.org/10.1186/s40104-017-0164-6

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M. O., Chupin, M., Benali, H., & Colliot, O. (2011). Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *NeuroImage*, *56*(2), 766–781. https://doi.org/10.1016/j.neuroimage.2010.06.013

deAndrés-Galiana, E. J., Fernández-Martínez, J. L., & Sonis, S. T. (2016). Design of biomedical robots for phenotype prediction problems. *Journal of Computational Biology*, *23*(8), 678–692. https://doi.org/10.1089/cmb.2016.0008

Edward, J., & Rosli, M. M. (2021). A systematic mapping study on ensemble-based classifier. In *2021 IEEE International Conference on Computing (ICOCO)* (pp. 43-48). IEEE Publishing. https://doi.org/10.1109/ICOCO53166.2021.9673563

Elbattah, M., Loughnane, C., Guérin, J.-L., Carette, R., Cilia, F., & Dequen, G. (2021). Variational autoencoder for image-based augmentation of eye-tracking data. *Journal of Imaging*, *7*(5), Article 83. https://doi.org/10.3390/jimaging7050083

Elkan, C. (2013). The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence* (Vol. 17, No. 1, pp. 973-978). Lawrence Erlbaum Associates Ltd.

Harimoorthy, K., & Thangavelu, M. (2021). Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. *Journal of Ambient Intelligence and Humanized Computing*, *12*(3), 3715–3723. https://doi.org/10.1007/s12652-019-01652-0

Japkowicz, N. (2000, June 28 – July 1). The class imbalance problem: Significance and strategies. In *Proceedings of the 2000 International Conference on Artificial Intelligence* (pp. 111-117). Las Vegas, NV, USA.

Jiang, Z., Ji, R., & Chang, K.-C. (2020). A machine learning integrated portfolio rebalance framework with risk-aversion adjustment. *Journal of Risk and Financial Management*, *13*(7), Article 155. https://doi.org/10.3390/jrfm13070155

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, *11*(1), 1-13. https://doi.org/10.1186/1472-6947-11-51

Krishnan, U., & Sangar, P. (2021). A rebalancing framework for classification of imbalanced medical appointment no-show data. *Journal of Data and Information Science*, *6*(1), 178–192. https://doi.org/doi:10.2478/jdis-2021-0011

Kuncheva, L. I. (2014). *Combining pattern classifiers*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118914564

Liang, C., Bian, Z., Lyu, W., Zeng, D., & Ma, J. (2018). A deep features-based radiomics model for breast lesion classification on FFDM. In *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)* (pp. 1-4). IEEE Publishing. https://doi.org/10.1109/NSSMIC.2018.8824722

Liu, Z., Cao, W., Gao, Z., Bian, J., Chen, H., Chang, Y., & Liu, T. Y. (2020). Self-paced ensemble for highly imbalanced massive data classification In *2020 IEEE 36th International Conference on Data Engineering (ICDE)* (pp. 841-852). IEEE Publishing. https://doi.org/10.1109/ICDE48307.2020.00078

Ma, T., Wu, L., Zhu, S., & Zhu, H. (2022). Multiclassification prediction of clay sensitivity using extreme gradient boosting based on imbalanced dataset. *Applied Sciences*, *12*(3), Article 1143. https://doi.org/10.3390/app12031143

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

Mohammed, R. A., Wong, K. W., Shiratuddin, M. F., & Wang, X. (2020). Pwidb: A framework for learning to classify imbalanced data streams with incremental data re-balancing technique. *Procedia Computer Science*, *176*, 818–827. https://doi.org/10.1016/j.procs.2020.09.077

Mohandes, M., Deriche, M., & Aliyu, S. O. (2018). Classifiers combination techniques: A comprehensive review. *IEEE Access*, *6*, 19626–19639. https://doi.org/10.1109/ACCESS.2018.2813079

Pes, B. (2019). Handling class imbalance in high-dimensional biomedical datasets. In *2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)* (pp. 150-155). IEEE Publishing. https://doi.org/10.1109/WETICE.2019.00040

Phoungphol, P., Zhang, Y., & Zhao, Y. (2012). Robust multiclass classification for learning from imbalanced biomedical data. *Tsinghua Science and Technology*, *17*(6), 619–628. https://doi.org/10.1109/TST.2012.6374363

Rahman, M. M., & Davis, D. N. (2013). Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing, 3*(2), Article 224. https://doi.org/10.7763/ijmlc.2013.v3.307

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, *10*(3), 1–21. https://doi.org/10.1371/journal.pone.0118432

Sandhan, T., & Choi, J. Y. (2014). Handling imbalanced datasets by partially guided hybrid sampling for pattern recognition. In *2014 22nd International Conference on Pattern Recognition* (pp. 1449-1453). IEEE Publishing. https://doi.org/10.1109/ICPR.2014.258

Shabaniyan, T., Parsaei, H., Aminsharifi, A., Movahedi, M. M., Jahromi, A. T., Pouyesh, S., & Parvin, H. (2019). An artificial intelligence-based clinical decision support system for large kidney stone treatment. *Australasian Physical and Engineering Sciences in Medicine*, *42*(3), 771–779. https://doi.org/10.1007/s13246-019-00780-3

Song, L., Lin, J., Wang, Z. J., & Wang, H. (2020). An end-to-end multi-task deep learning framework for skin lesion analysis. *IEEE Journal of Biomedical and Health Informatics*, *24*(10), 2912–2921. https://doi.org/10.1109/JBHI.2020.2973614

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

Tantithamthavorn, C., Hassan, A. E., & Matsumoto, K. (2020). The impact of class rebalancing techniques on the performance and interpretation of defect prediction models. *IEEE Transactions on Software Engineering*, *46*(11), 1200–1219. https://doi.org/10.1109/TSE.2018.2876537

Turlapati, V. P. K., & Prusty, M. R. (2020). Outlier-SMOTE: A refined oversampling technique for improved detection of COVID-19. *Intelligence-Based Medicine*, *3–4*, Article 100023. https://doi.org/10.1016/j.ibmed.2020.100023

Utami, I. T., Sartono, B., & Sadik, K. (2014). Comparison of single and ensemble classifiers of support vector machine and classification tree. *Journal of Mathematical Sciences and Applications*, *2*(2), 17–20. https://doi.org/10.12691/jmsa-2-2-1

Valentini, G., & Dietterich, T. G. (2004). Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods. *Journal of Machine Learning Research*, *5*, 725–775.

Yao, J. R., & Chen, J. R. (2019). A new hybrid support vector machine ensemble classification model for credit scoring. *Journal of Information Technology Research*, *12*(1), 77–88. https://doi.org/10.4018/JITR.2019010106

Zhao, Y., Wong, Z. S. Y., & Tsui, K. L. (2018). A framework of rebalancing imbalanced healthcare data for rare events' classification: A case of look-alike sound-alike mix-up incident detection. *Journal of Healthcare Engineering*, *2018*(2010), Article 6275435. https://doi.org/10.1155/2018/6275435

Zhu, R., Guo, Y., & Xue, J.-H. (2020). Adjusting the imbalance ratio by the dimensionality of imbalanced data. *Pattern Recognition Letters*, *133*, 217–223. https://doi.org/10.1016/j.patrec.2020.03.004

Zhu, W., Oh, B. S., Huang, W., Lin, Z., Pan, Y., & Zhou, J. (2015). Hybrid classifiers ensemble with an undersampling scheme for liver tumor segmentation. In *2015 10th International Conference on Information, Communications and Signal Processing (ICICS)* (pp. 1-4). IEEE Publishing. https://doi.org/10.1109/ICICS.2015.7459850