IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 14, No. 5, October 2025, pp. 3982~3993
ISSN: 2252-8938, DOI: 10.11591/ijai.v14.i5.pp3982-3993

A hybrid model for handling the imbalanced multiclass classification problem

Esra'a Alshdaifat[1], Fairouz Hussein[2], Ala'a Al-shdaifat[1], Malak Al-Hassan[3], Enshirah Altarawneh[4]
[1]Department of Information Technology, Faculty of Prince Al-Hussein Bin Abdallah II for Information Technology, The Hashemite University, Zarqa, Jordan
[2]Department of Computer Information Systems, Faculty of Prince Al-Hussein Bin Abdallah II for Information Technology, The Hashemite University, Zarqa, Jordan
[3]King Abdullah II School of Information Technology, The University of Jordan, Amman, Jordan
[4]Department of Computer Engineering, Faculty of Engineering, The Hashemite University, Zarqa, Jordan

Article Info

Article history:

Received Oct 31, 2024
Revised Jul 5, 2025
Accepted Aug 6, 2025

Keywords:

Boosting
Bootstrap
Imbalanced data
Multiclass classification
One-vs-one
SMOTE

ABSTRACT

Data in many application domains is imbalanced. In machine learning, addressing imbalanced data is crucial to prevent bias towards the dominant class label and ensure that prediction models can learn and predict the minority class proficiently. This paper proposes a hybrid imbalanced classification model (HICD) to address the multiclass imbalanced data problem. The primary idea is to combine effective methods to construct a classification model that can handle multiclass imbalanced data effectively. Four methods are employed: an oversampling method to balance the data, a decomposition method to convert the multiclass problem into a set of binary problems, ensemble classification to integrate base classifiers to improve prediction, and a boosting method to encourage the classifier to pay more attention to misclassified samples. To evaluate the proposed model, seventeen imbalanced datasets from various application domains, featuring different numbers of classes, instances, features, and imbalance ratios, are assessed. The experimental results and statistical significance tests demonstrate that the proposed hybrid model significantly outperforms the standard one-vs-one (OVO) approach and the OVO combined with oversampling technique (SMOTE), both considered state-of-the-art for addressing imbalanced multiclass datasets, in terms of F1-score.

Corresponding Author:

Esra'a Alshdaifat
Department of Information Technology
Faculty of Prince Al-Hussein Bin Abdallah II for Information Technology, The Hashemite University
P.O. Box 330127, Zarqa 13133, Jordan
Email: esraa@hu.edu.jo

1. INTRODUCTION

In several real-world problems, such as disease identification, text classification, network intrusion detection, and spam filtering, imbalanced data is common. Where the frequency of class labels in the dataset is unequal, in other words, one or more classes are underrepresented, in contrast, the remaining classes are highly represented in the dataset. The class represented by a significantly larger number of observations relative to other classes, in the dataset, is referred to as the "majority class". While the class that is represented by a noticeably smaller number of observations relative to other classes is referred to as the "minority class". Two main imbalance problems can be identified: binary imbalanced problem, where the dataset contains only two classes (the majority and minority class), and multiclass imbalanced problem, which includes more than two classes, with one or more classes represented by fewer instances.

Using the standard machine learning algorithms as they are on an imbalanced dataset will result in majority class label bias, and the accuracy of the produced model will not be representative of its actual usefulness. To make this problem clear, imagine a medical diagnosis data set having two classes: i) majority class (negative), which forms 95% of samples, and ii) minority (positive) class, which forms 5% of samples. Creating a classification model that constantly outputs the majority class, gives an accuracy rate of 95%. In this scenario, the samples of the minority class were neglected by the classification algorithm, and the obtained accuracy score is considered misleading. Note here that greater importance is often given to the underrepresented class. For instance, in the previous medical diagnosis problem, the minority class is the "positive" samples, which are rare but essential to be detected precisely. The same issue occurs in multiclass imbalanced problems; however, it is more challenging. Considering a heart disease dataset, where patients are categorized into five classes based on the severity of heart disease, which range from class 0 (no disease) to classes one to four (severe diseases). The no disease and non-severe disease classes are the dominant classes, whereas classes that represent more severe cases are represented by fewer samples. Training a classifier on this dataset will be effective in predicting no or mild disease classes, but it might not be able to detect patients belonging to more severe classes.

From the foregoing, handling imbalanced datasets is considered a challenging and well-known problem in the machine learning field. Consequently, much research work has been conducted by numerous researchers to tackle this problem. The work in addressing the imbalanced data problem can be categorized into three main categories [1]: i) data-level methods, ii) algorithmic-level methods, and iii) hybrid methods. In data-level methods, balancing the data is performed by augmenting the minority class observations or reducing the majority class observations, which are known as over-sampling and undersampling methods. Concerning the algorithmic-level methods, such methods involve modifying existing algorithms or proposing a structure for new algorithms to address the imbalanced data problem. With respect to hybrid methods, a combination of data-level and algorithmic-level methods is employed to handle the imbalanced data.

The solution proposed in this paper for handling the imbalanced data problem belongs to the hybrid methods category. More specifically, four methods are combined to tackle the imbalanced data and obtain an effective classification model. The first method is a data-level method: the well-known synthetic minority oversampling technique (SMOTE) is utilized [2]. The second method is an ensemble method, where a collection of classifiers is utilized to enhance classification effectiveness. The third method is a decomposition method, in which a multiclass classification problem is decomposed into a number of binary sub-problems, and each classifier focuses only on two classes; thus, better classification effectiveness can be obtained. The fourth method is a boosting method, which identifies the low-performance base classifiers and forces them to focus on misclassified instances using a bootstrap technique. The idea is that integrating four effective methods for handling imbalanced multiclass classification can result in a high-performance hybrid model. Further information about the proposed model is provided in section 3.

The rest of this paper is structured in the following sections: section 2 provides an overview of the methods used to handle imbalanced datasets. Section 3 explains the generation and use of the suggested hybrid imbalanced multiclass classification model. Section 4 presents a general description of the evaluation datasets. Section 5 covers the experimental setup and reports the produced results. Section 6 summarizes the paper and provides some directions for future work.

2. LITERATURE REVIEW

In this section, an overview of the methods used to handle imbalanced datasets is presented. As mentioned earlier, imbalanced datasets can be handled using three primary methods [1]: i) data-level methods, ii) algorithmic-level methods, and iii) hybrid methods. Commencing with the data-level methods, which are used to balance the data during the preprocessing phase. These methods can be divided into two groups: oversampling and undersampling methods. In oversampling, the class imbalance is addressed by increasing the number of minority class samples. This can be achieved by either duplicating existing minority class instances randomly or by generating new synthetic samples. The first approach involves repeating some instances, which is straightforward but may cause overfitting. The second approach applies interpolation between minority class observations to generate new observations, such as using the SMOTE [2], the result here is more diverse synthetic samples [3]. SMOTE is considered the most widely used oversampling method and has broad applications [4]. Many researchers applied it to imbalanced data problems and reported that the model performance improved significantly [5], [6]. On the other hand, some researchers argue that the resulting synthetic samples may not accurately reflect the original data, and they referred to the new samples as "unrealistic samples", arguing that this can degrade classifier accuracy [7]. Adaptive synthetic (ADASYN) sampling approach for imbalanced learning also creates synthetic examples, but it adopts a more adaptive way compared to traditional SMOTE [3].

Regarding the undersampling methods, samples are removed from the majority class until the dataset becomes balanced. This is done to avoid bias in classification models toward the majority class [8]. Random undersampling (RUS), is considered one of the simplest and most common undersampling methods, in which samples from majority classes are removed randomly. However, this leads to a loss of valuable information that could impact the performance of the resulting model [9]. Consequently, other methods emerged and attempted to remove samples from the majority classes based on some defined criteria, such as the radial-based undersampling algorithm [10].

With respect to the algorithmic-level methods, which are also known as "internal approaches", the data imbalance problem is handled by creating or improving existing classification algorithms [4]. These methods include threshold adjustments, one-class learning, cost-sensitive learning, and ensemble-based techniques [4], [11]–[13]. In the threshold adjustment method, classifiers often provide probabilities that refer to which class an observation belongs, which can be used to adjust thresholds and refine class assignments [11]. Cost-sensitive learning assigns greater misclassification costs to minority class samples to encourage the classifier to pay more attention to underrepresented samples [4]. One class classification focuses on the minority class and learning its characteristics to differentiate it from the other data [11]. Ensemble classifiers aim to enhance the performance of classification tasks by combining predictions from a set of base classifiers [14]. Common ensemble methods include bagging and boosting [14]. Using ensembles of classifiers has become a popular method for addressing class imbalance in machine learning [11], [12]. Some research works focused on simplifying and converting the single multiclass problem into many binary problems using specific decomposition techniques, such as one-vs-one (OVO), one-vs-all (OVA), and the binary tree method [15]. The idea here is to focus on one or two classes instead of creating a model that differentiates between several classes.

Some researchers focused their research on combining data-level methods and algorithmic-level methods to generate more powerful models to handle the imbalance class problem, these methods are referred to as hybrid methods [16]. It is important to note that hybrid models can be differentiated according to: i) the adopted data and algorithm methods, and ii) whether the addressed classification problem is binary or multiclass. Most research work related to the generation of hybrid imbalanced models has been conducted on binary imbalanced problems. Commencing with the binary hybrid model proposed by Sun *et al*. [17], in which the bagging ensemble method is combined with SMOTE. Shi *et al*. [18] integrated a novel density-based sampling technique with the ensemble approach to construct a binary hybrid imbalanced classification model (HICD). HICD partitions the data space into five areas according to data density, and then the data is sampled from these areas. Once the data is sampled, the ensemble model is generated. While the model proposed by Theephoowiang and Hanskunatai [19] splits the data into four different groups according to the overlapping and non-overlapping concept between the majority and minority classes instances, the data categories are then used to form five datasets, which are resampled using different SMOTEs. The sampled datasets are then used to generate the classification models using different single and ensemble algorithms. Shan and Chung [20] coupled data-level techniques and loss function to generate the desired hybrid model. The suggested model begins with dividing samples based on their effect on imbalanced data classification into several categories, thus appropriate samples can be selected for sampling. A loss function is then proposed, relying on sample difficulty.

Multiclass imbalanced classification problem is considered challenging research due to the complexities caused by multiple classes [21]. Several researchers tried to combine the ensemble methods, such as bagging or boosting, with oversampling or undersampling techniques to address the multiclass imbalanced problem [22]. More recent work on multiclass imbalanced hybrid model generation is focused on proposing unique data-level methods and combining them with the ensemble methods or integrating the state-of-the-art sampling methods with a novel algorithmic-level method. The work proposed by Hartono *et al*. [23] introduced a generalization potential and learning difficulty-based hybrid sampling (GDHS) method as a data-level method and combined it with the gradient boosting decision tree (DT) ensemble model. In GDHS, minority class representation is improved by applying intelligent oversampling, and the majority classes are cleaned to minimize noise and overlap. Some researchers tried to combine OVO or OVA with oversampling methods and ensemble classification or deep learning, such as the work proposed in [24], [25]. Salehi and Khedmati [21] suggested a hybrid cluster-based oversampling and undersampling (HCBOU) technique, which clusters classes into majority and minority groups to guide the sampling process. HCBOU preserves the class structure and produces convenient synthetic samples. The novel HCBOU is integrated with OVO and OVA classification decomposition methods.

The work presented in this paper is directed at generating a hybrid imbalanced multiclass classification model. The core idea is to integrate four well-known powerful methods for handling imbalanced data problem, to construct a high-performance hybrid model. More specifically, the utilized methods are:

– SMOTE method, in which the minority class is oversampled to balance the data and improve generalization.
– OVO method, in which a multiclass dataset is mapped into a number of binary datasets, and a classifier is generated for each. This simplification can produce better classification effectiveness.
– Ensemble method, in which several classifiers are joined to enhance classification effectiveness. Note here that the binary classifiers generated using OVO decomposition are considered a form of ensemble. Moreover, an ensemble of classifiers that can be used as a base classifier for each class pair is a form of ensemble, and both forms are considered in the work presented in this paper.
– Boosting method, in which each base classifier within the ensemble is evaluated, and those with lower performance are boosted to focus more on the samples they misclassified.

3. THE HYBRID IMBALANCED MULTICLASS CLASSIFICATION MODEL

This section illustrates the construction and use of the hybrid imbalanced multiclass classification model. Again, the fundamental idea is to merge: i) oversampling, ii) ensemble, iii) decomposition, and iv) boosting methods to construct an effective classification model for imbalanced multiclass classification problems. Figure 1 presents an example of the desired model generation process for a dataset including four class labels. The process begins with applying the SMOTE to balance the data. Next, the multiclass dataset is decomposed into multiple binary datasets using the OVO approach. An initial set of base classifiers is then trained and evaluated. Based on the evaluation results, each base classifier is either boosted or not, and afterward retrained on the entire corresponding binary dataset to avoid any data loss. As a result, a set of balanced and boosted base classifiers is generated, collectively forming the final desired hybrid model. Although the model generation process involves several stages, it is performed only once.

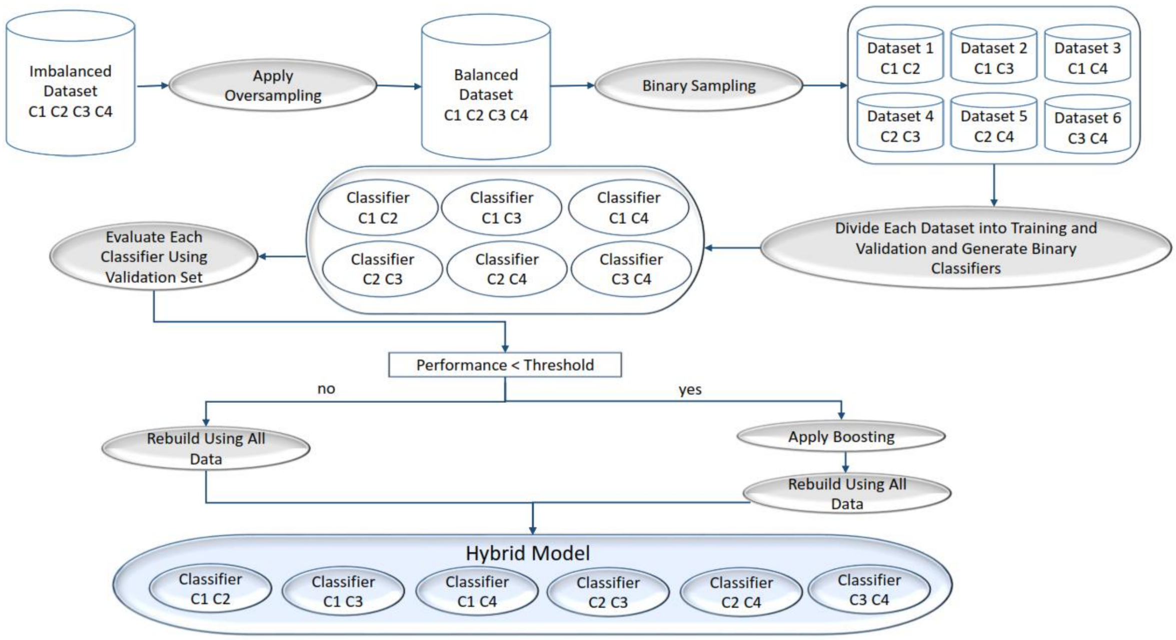

Figure 1. The generation process of the hybrid imbalanced multiclass classification model

The detailed process of model construction is explained in Algorithm 1. The algorithm has five inputs: i) the input dataset *D*, ii) the set of classes *C*, iii) the SMOTE that will be used to balance the data *O*, iv) the classification algorithm *Algo* that will be utilized to construct the base classifiers, and v) the performance threshold *acc_threshold* that will be adopted to spot the classifiers that need to be boosted. The algorithm begins by applying the SMOTE to the dataset to produce the balanced resampled data *Resampled_D* (line 9). Then, all possible combinations of size two classes featured in the dataset will be found (line 10). The algorithm then loops through the set of class combinations, and on each iteration, it finds a set of examples D_i in *D* that features C_i (line 11). Then it divides D_i into training and validation sets, thus a classifier can be built and evaluated to generate an accuracy score acc_i (lines 14 and 15). The next step is to identify weak classifiers by comparing the evaluated accuracy score with the accuracy threshold (line 16). If the accuracy score is under the predefined threshold, the bootstrap method is applied to the misclassified data, and the result is added to the D_i data and used to rebuild the boosted base classifier *boosted_classifier*$_i$, which is then added to the set of base classifiers forming the hybrid model (lines 16 to 20). While if the accuracy score is above the predefined threshold, then the base classifier is reconstructed using the training data D_i without applying boosting and then added to the set of base classifiers forming the hybrid model (lines 22 and 23). The hybrid classification model is the output of the algorithm, which consists of a set of binary balanced base classifiers.

Algorithm 1. Hybrid imbalanced multiclass classification model construction

```
1: INPUT
2: D: the input dataset
3. C: the unique classes in D
4. O: the oversampling technique
5: Algo: the classification algorithm
6: acc_threshold: accuracy threshold
7: OUTPUT
8: The generated hybrid classification model
9: Resampled_D = Apply O on D
10: C_combinations = Find all sets of size 2 combinations in C
11: for i =1 to j =|C_combinations| do
12:  D_i = Find set of examples in D that features C_i
13:  T_i, V_i = divide D_i into training and validation sets
14:  classifier_i = Use Algo to construct base classifier classifier_i using training set T_i
15:  acc_i = use V_i to evaluate classifier_i
16:  if (acc_i< acc_threshold)
17:  boosted_misclassified_i = apply bootstrap on misclassified data
18:  boosted_D_i = D_i ∪ boosted_misclassified_i
19:  boosted_classifier_i = Use Algo to construct base classifier using boosted_D_i
20:  hybrid_model = hybrid_model ∪ boosted_classifier_i
21:  else
22:  classifier_i = Use Algo to construct base classifier C_i using training set D_i
23.  hybrid_model = hybrid_model ∪ classifier_i
24:  end if
25: end for
```

When using the generated hybrid model for prediction, a majority voting approach is adopted to aggregate the predictions from the member binary classifiers. More particularly, to classify a new unseen sample, all the individual binary classifiers in the generated hybrid classification model are utilized to classify the sample, and the class label that receives the majority of votes is considered the final output and is assigned to the unseen sample. Hence, the well-known SMOTE method is utilized, and the adopted decomposition method is the OVO; we will refer to the hybrid model as Boosted-OVO&SMOTE throughout the rest of the paper.

For evaluating the resulting model, the accuracy, precision, recall, and F1-score are considered:

– Accuracy: the ratio of correctly predicted observations to all observations in a given test set [26].

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

– Precision: the ratio of observations correctly predicted as positive to all observations predicted as positive [26].

$$precision = \frac{TP}{TP+FP} \tag{2}$$

– Recall: the ratio of observations correctly predicted as positive to all actual positive observations [26].

$$recall = \frac{TP}{TP+FN} \tag{3}$$

– F1-score: it represents a combination of the precision and recall scores [26].

$$F1-score = \frac{2*Precision*Recall}{Precision+Recall} \tag{4}$$

Here, TP denotes the true positive, TN denotes the true negative, FP denotes the false positive, and FN denotes the false negative records. Because the datasets taken into consideration in this study are multiclass datasets, macro scores are utilized.

4. DATASETS

This section provides a summary of the main attributes of the datasets used to assess the proposed hybrid model. Seventeen imbalanced datasets from various disciplines, each with a different number of observations, classes and attributes, all sourced from the University of California Irvine (UCI) Machine Learning Repository [27]. Table 1 outlines the key features of these datasets. Because the research presented in this paper focuses on imbalanced multiclass classification problems, the datasets include a range of class distribution rates.

Table 1. The description of the experimental datasets

Dataset	# of classes	# of features	# of instances	Distribution of classes ratio	Domain
Abalone	3	8	4177	1407/2406/364 (Ratio =33.7: 57.6: 8.7)	Biology
Contraceptive method	3	9	1473	415/227/831 (Ratio =28.17: 15.41: 56.42)	Health and medicine
Hayes-Roth	3	4	160	65/64/31 (Ratio =40.63: 40.00: 19.38)	Social science
Post-operative	3	8	90	2/24/64 (Ratio =2.22: 26.67: 71.11)	Health and medicine
Thyroid	3	5	215	150/35/30 (Ratio =69.7: 16.3: 14.0)	Health and medicine
Vertebral	3	6	310	60/150/100 (Ratio =19.35:48.39:32.26)	Health and medicine
Vehicle	4	18	846	199/217/218/212 (Ratio =23.52: 25.66: 25.79: 25.03)	Automotive
Car	4	6	1728	1210/384/65/69 (Ratio =70.0: 22.2: 3.8: 4.0)	Automotive
Heart (Cleveland)	5	13	297	160/54/35/35/13 (Ratio =53.9: 18.2: 11.8: 11.8: 4.4)	Health and medicine
Nursery	5	8	12960	4320/2/328/4266/4044 (Ratio =33.3:0.015:2.5:32.9:31.2)	Social science
Page blocks	5	10	5473	4913/329/28/88/115 (Ratio =89.8:6.0:0.5:1.6:2.1)	Computer science
Dermatology	6	34	366	112/61/72/49/52/20 (Ratio =30.6:16.7:19.7:13.4:14.2:5.5)	Health and Medicine
Dry bean	7	16	13611	2027/1322/522/1630/1928/2636/3546 (Ratio =14.9:9.7:3.8:12.0:14.2:19.3:26.0)	Biology
Glass	7	9	214	70/17/0/76/13/9/29 (Ratio =32.7:7.9:0.0:35.5:6.1:4.2:13.6)	Physics and chemistry
E. coli	8	7	336	143/77/52/35/20/5/2/2 (Ratio =42.5:22.9:15.4:10.4:5.9:1.5:0.6:0.6)	Biology
Pen digits	10	16	10992	1143/1143/1144/1055/1144/1055/1056/1142/1055/1055 (Ratio =10.4:10.4:10.4:9.6:10.4:9.6:9.6:10.4:9.6:9.6)	Computer science
Yeast	10	8	1484	244/429/ 463/44/35/51/163/30/ 20/5 (Ratio =16.4:28.9:31.2:3.0:2.4:3.4:11.0:2.0:1.3:0.3)	Biology

5. EXPERIMENTS AND ANALYSIS

This section discusses the experimental setup and reports the obtained results. For building the individual classifiers, three algorithms were employed: i) DT, ii) support vector machine (SVM), and iii) random forest (RF). These algorithms were chosen because of: i) their different learning behaviors, which enable comprehensive evaluation of the effectiveness of the suggested hybrid model to be conducted, and ii) their popularity and reported performance in prediction. DT is well-known for its simplicity and interpretability, SVM is effective in high-dimensional spaces, and RF, as an ensemble classification method, is recognized for improving classification effectiveness. To ensure precise results, ten-fold cross validation (TCV) was employed for all the experiments reported in this paper. The evaluation measures included accuracy, precision, recall, and F1-score. To simplify the analysis, the results will be discussed based on the F1-score because: i) it combines two measures; precision and recall, and ii) it reflects precise performance for imbalanced datasets. With respect to SMOTE, the k-neighbors parameter is set to one because some evaluation datasets include only two samples for the minority class. The SVM classifier employed the radial basis function (RBF) kernel. Fifty classifiers were constructed as base classifiers for the RF classifier. Each dataset is evaluated using three methods coupled with three classification algorithms. More specifically, for each classification algorithm, the methods are: i) OVO with one of the base classifiers (OVO), ii) OVO and SMOTE (OVO&SMOTE), and iii) OVO coupled with SMOTE and bootstrap boosting (Boosted-OVO&SMOTE). As noted earlier, a threshold value is utilized to spot the classifiers that should be boosted; several experiments were conducted to identify the best threshold value for each dataset and classification algorithm. Table 2 presents the adopted threshold values for each considered evaluation dataset and classifier. The produced results are presented and discussed in the next sub-sections.

Table 2. The adopted boosting threshold values

Dataset	Best boosting threshold value		
	DT boosting threshold	SVM boosting threshold	RF Boosting threshold
Abalone	0.75	0.80	0.95
Contraceptive	0.70	0.70	0.75
Hayes Roth	0.85	0.90	0.70
Post-operative	0.85	0.72	0.65
Thyroid	0.99	0.95	0.95
Vertebral	0.95	0.70	0.95
Vehicle	0.90	0.99	0.75
Car	0.95	0.99	0.99
Heart	0.75	0.99	0.80
Nursery	0.95	0.95	0.95
Page blocks	0.95	0.90	0.95
Dermatology	0.94	0.95	0.95
Dry bean	0.90	0.95	0.99
Glass	0.89	0.70	0.95
E. coli	0.99	0.90	0.90
Pen digits	0.99	0.99	0.99
Yeast	0.90	0.60	0.80

5.1. Results obtained from using the DT classifier to construct the hybrid model

In this section, the results produced from using the DT classifier to generate the desired hybrid model are presented and discussed. The results are tabulated in Table 3, and the best results are highlighted in bold font. Commencing with comparing the performance of OVO and OVO SMOTE models, from the table, it is clear that combining SMOTE and OVO outperforms using OVO alone. The same observation is noticed when comparing the results obtained from using the Boosted-OVO&SMOTE hybrid model and the OVO model. Thus, combining OVO and SMOTE to generate a hybrid model improved the classification effectiveness. Regarding comparing the proposed hybrid model (Boosted-OVO&SMOTE) with OVO&SMOTE, it is obvious that the hybrid model outperforms the OVO&SMOTE model. More specifically, Boosted-OVO&SMOTE generated the best results for all the considered datasets. However, for six datasets, the same results were obtained from using OVO&SMOTE. Consequently, boosting the relatively low-performance classifiers resulted in improving the classification effectiveness.

Table 3. Results obtained from using the DT classifier as the base classifier

Dataset	OVO				OVO&SMOTE				Boosted-OVO&SMOTE			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Abalone	0.668	0.571	0.589	0.575	0.800	0.799	0.800	0.799	0.800	0.799	0.800	0.799
Contraceptive	0.464	0.449	0.441	0.439	0.569	0.569	0.570	0.565	0.574	0.573	0.574	0.571
Hayes Roth	0.825	0.864	0.855	0.853	0.837	0.844	0.827	0.814	0.841	0.852	0.843	0.826
Post-operative	0.643	0.536	0.527	0.521	0.797	0.804	0.795	0.789	0.823	0.842	0.825	0.816
Thyroid	0.944	0.942	0.923	0.922	0.964	0.964	0.963	0.963	0.976	0.977	0.974	0.975
Vertebral	0.806	0.770	0.756	0.752	0.833	0.841	0.830	0.828	0.849	0.846	0.846	0.843
Vehicle	0.704	0.723	0.706	0.709	0.716	0.720	0.714	0.711	0.746	0.750	0.746	0.743
Car	0.859	0.863	0.816	0.790	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998
Heart	0.459	0.279	0.270	0.268	0.791	0.798	0.788	0.786	0.791	0.798	0.788	0.786
Nursery	0.848	0.866	0.835	0.812	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
Page blocks	0.959	0.817	0.809	0.799	0.991	0.991	0.991	0.991	0.991	0.991	0.991	0.991
Dermatology	0.944	0.948	0.948	0.942	0.974	0.974	0.969	0.970	0.976	0.975	0.971	0.972
Dry bean	0.703	0.783	0.764	0.733	0.944	0.944	0.944	0.944	0.944	0.944	0.944	0.944
Glass	0.711	0.691	0.716	0.683	0.866	0.865	0.875	0.858	0.890	0.894	0.895	0.885
E. coli	0.809	0.704	0.689	0.672	0.940	0.939	0.940	0.937	0.948	0.947	0.951	0.947
Pen digits	0.965	0.966	0.965	0.965	0.969	0.969	0.969	0.969	0.971	0.971	0.971	0.971
Yeast	0.487	0.449	0.458	0.442	0.843	0.848	0.842	0.843	0.849	0.853	0.849	0.849

5.2. Results obtained from using the support vector machine classifier to construct the hybrid model

In this section, the results achieved from using the SVM classifier to generate the desired hybrid model are presented and discussed. The results are tabulated in Table 4. Again, combining SMOTE with OVO outperforms using OVO alone. Regarding comparing the Boosted-OVO&SMOTE and OVO&SMOTE, the boosted approach generated the same or better results for all the considered datasets. More specifically, Boosted-OVO&SMOTE produced better results for eight of the seventeen considered datasets (abalone, contraceptive, Hayes Roth, thyroid, car, heart, glass, and E. coli), for the nine remaining datasets the same results were produced by the OVO&SMOTE model. Note here that the observations from using the DT and SVM classifiers are harmonic.

Table 4. Results obtained from using the SVM classifier as the base classifier

Dataset	OVO				OVO&SMOTE				Boosted-OVO&SMOTE			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Abalone	0.748	0.516	0.522	0.510	0.716	0.722	0.716	0.718	0.723	0.722	0.723	0.720
Contraceptive	0.489	0.474	0.472	0.470	0.517	0.537	0.518	0.509	0.516	0.517	0.515	0.510
Hayes Roth	0.550	0.632	0.573	0.569	0.652	0.657	0.665	0.637	0.738	0.766	0.749	0.728
Post-operative	0.688	0.287	0.417	0.339	0.798	0.815	0.802	0.791	0.798	0.815	0.802	0.791
Thyroid	0.953	0.947	0.892	0.892	0.973	0.975	0.973	0.973	0.980	0.982	0.979	0.979
Vertebral	0.816	0.801	0.781	0.781	0.793	0.795	0.794	0.787	0.793	0.795	0.794	0.787
Vehicle	0.751	0.739	0.753	0.738	0.767	0.755	0.769	0.754	0.763	0.763	0.763	0.754
Car	0.929	0.929	0.858	0.859	0.989	0.990	0.989	0.989	0.990	0.990	0.990	0.990
Heart	0.572	0.233	0.276	0.251	0.720	0.723	0.719	0.705	0.746	0.755	0.747	0.743
Nursery	0.907	0.880	0.848	0.847	0.989	0.989	0.989	0.989	0.989	0.989	0.989	0.989
Page blocks	0.939	0.688	0.526	0.570	0.937	0.938	0.937	0.937	0.937	0.938	0.937	0.937
Dermatology	0.975	0.975	0.971	0.972	0.986	0.985	0.988	0.986	0.986	0.985	0.988	0.986
Dry bean	0.895	0.925	0.911	0.906	0.940	0.941	0.940	0.940	0.940	0.941	0.940	0.940
Glass	0.682	0.517	0.540	0.514	0.765	0.774	0.771	0.756	0.785	0.798	0.786	0.778
E. coli	0.860	0.807	0.767	0.770	0.899	0.904	0.900	0.897	0.909	0.914	0.910	0.907
Pen digits	0.994	0.994	0.994	0.994	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995
Yeast	0.598	0.579	0.538	0.541	0.668	0.699	0.668	0.671	0.668	0.699	0.668	0.671

5.3. Results obtained from using the random forest ensemble to construct the hybrid model

This section illustrates and describes the experimental results produced when the RF ensemble classifier was utilized as the base classifier to generate the suggested hybrid model. The results are tabulated in Table 5. Like the case of DT and SVM classifiers, Boosted-OVO&SMOTE using RF as base classifiers produced the best F1-score for most datasets. Moreover, employing RF as the base classifier resulted in further performance improvements. More specifically, it achieved the highest F1-score for sixteen out of the seventeen datasets considered in the investigation, although for four of those datasets, the OVO&SMOTE model achieved the same score.

Now, to show that Boosted-OVO&SMOTE significantly outperforms OVO and OVO&SMOTE hybrid models the Friedman statistical significance test [28] was applied. According to Friedman test statistics, there is a significant difference in performance among the hybrid models ($X^2(2) = 4.5000$, $p = 0.00000$). Accordingly, the Nemenyi post-hoc test [29] was employed to identify the superior hybrid model. Figure 2 displays the output of the Nemenyi post-hoc test. Note here that to consider one model significantly exceeds another, the difference between their calculated average ranks should be greater than or equal to a critical difference (CD). From the figure, both Boosted-OVO&SMOTE and OVO&SMOTE models perform better than the OVO model, further, the Boosted-OVO&SMOTE hybrid model significantly outperforms the OVO&SMOTE hybrid model.

Table 5. Results obtained from using the RF ensemble as the base classifier

Dataset	OVO				OVO&SMOTE				Boosted-OVO&SMOTE			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Abalone	0.757	0.598	0.721	0.614	0.869	0.868	0.869	0.868	0.871	0.870	0.871	0.870
Contraceptive	0.511	0.484	0.496	0.485	0.621	0.619	0.621	0.618	0.629	0.627	0.628	0.626
Hayes Roth	0.806	0.853	0.839	0.836	0.837	0.844	0.819	0.810	0.847	0.864	0.847	0.827
Post-operative	0.673	0.480	0.556	0.506	0.818	0.830	0.824	0.815	0.824	0.843	0.831	0.821
Thyroid	0.963	0.975	0.932	0.942	0.989	0.990	0.987	0.988	0.989	0.990	0.987	0.988
Vertebral	0.842	0.812	0.790	0.789	0.907	0.908	0.909	0.906	0.911	0.913	0.912	0.909
Vehicle	0.766	0.764	0.768	0.762	0.752	0.743	0.754	0.741	0.772	0.763	0.772	0.763
Car	0.858	0.792	0.766	0.743	0.996	0.996	0.996	0.996	0.998	0.998	0.998	0.998
Heart	0.556	0.241	0.262	0.242	0.883	0.886	0.888	0.883	0.883	0.886	0.888	0.883
Nursery	0.818	0.846	0.786	0.774	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998
Page blocks	0.968	0.890	0.825	0.837	0.994	0.994	0.994	0.994	0.994	0.994	0.994	0.994
Dermatology	0.978	0.979	0.974	0.975	0.989	0.989	0.991	0.989	0.991	0.992	0.991	0.991
Dry bean	0.730	0.783	0.782	0.751	0.957	0.957	0.957	0.957	0.958	0.958	0.958	0.958
Glass	0.791	0.700	0.735	0.706	0.923	0.927	0.930	0.922	0.925	0.926	0.930	0.923
E. coli	0.866	0.776	0.759	0.753	0.960	0.959	0.960	0.958	0.962	0.962	0.963	0.960
Pen digits	0.988	0.988	0.988	0.988	0.990	0.990	0.990	0.990	0.994	0.994	0.994	0.994
Yeast	0.588	0.529	0.499	0.500	0.889	0.890	0.889	0.888	0.890	0.891	0.890	0.889

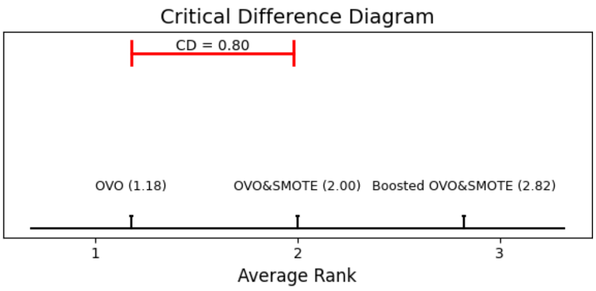

Figure 2. The result of the Nemenyi post-hoc test for comparing OVO model, OVO&SMOTE model, and Boosted-OVO&SMOTE model

5.4. Comparing the performance of the classifiers utilized to generate the hybrid model

This section presents a comparison among the base classifiers used to generate the desired hybrid model. Figure 3 displays the performance comparison in terms of F1-score for the three considered classifiers: i) DT, ii) SVM, and iii) RF classifiers. From the figure, it is obvious that the performance of the RF hybrid model outperforms DT and SVM hybrid models for the most considered datasets. More specifically, the RF hybrid model achieved the best F1-score for fifteen datasets, while the DT hybrid model generated the best F1-score for two datasets (car and nursery). Note that the same result was obtained for the car dataset using the DT and RF hybrid models. For only one dataset (pen digits), the SVM hybrid model produced the best F1-score. Therefore, the adopted classifier can significantly influence the overall effectiveness of the hybrid model. To conduct a precise comparison, the Friedman test was adopted and reported a significant difference among the considered models ($X^2(2) = 21.5224$, $p = 0.00002$). Therefore, the Nemenyi post-hoc test was employed to highlight the superior hybrid model. Figure 4 summarizes the output of the Nemenyi post-hoc test. From the figure, the RF hybrid model significantly outperforms both DT and SVM hybrid models. In addition, no significant difference between the DT hybrid model and the SVM hybrid model (connected models in the figure indicate no significant difference). In summary, utilizing RF to generate the desired hybrid model provides clear evidence that adopting ensemble classification improves the classification effectiveness for the hybrid model.

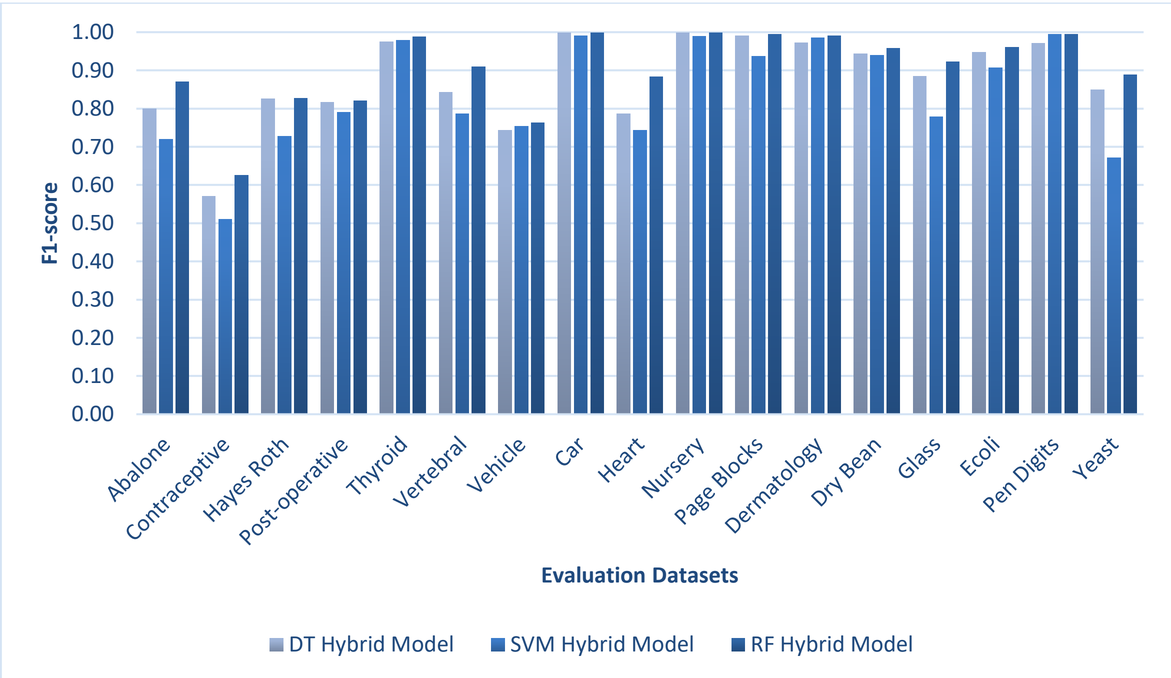

Figure 3. Comparing the performance of the three considered classifiers utilized to generate the hybrid model

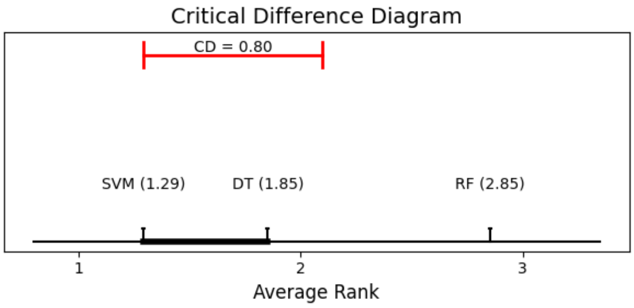

Figure 4. The result of the Nemenyi post-hoc test for comparing: RF hybrid model, DT hybrid model, and SVM hybrid model

6. CONCLUSION

In this paper, a novel solution to the well-known imbalanced multiclass classification problem belonging to the hybrid methods category is presented and illustrated. The primary idea is to combine four powerful methods for handling imbalanced multiclass classification to construct a high-performance hybrid model. The examined methods include: the well-known SMOTE data-level method, the decomposition method, the ensemble method and the boosting method. Regarding the ensemble and decomposition methods, these were achieved through the OVO approach, which involves decomposing the multiclass problem into multiple binary problems and constructing a tailored classifier for each binary problem. Concerning the boosting method, the idea was to identify the less effective classifiers and boost them using the bootstrap method. Consequently, our hybrid model is referred to as Boosted-OVO&SMOTE. According to the findings, the Boosted-OVO&SMOTE hybrid model significantly outperforms the conventional OVO model. Moreover, the suggested model improved classification effectiveness or preserved the same performance when compared with the OVO&SMOTE model, this indicates that the model is effective in spotting the classifiers that require boosting. In other words, the suggested model will produce better results when the binary classifiers within the OVO include relatively "low-performance" classifiers. Moreover, utilizing the RF ensemble classifier as a base classifier significantly enhances the overall performance compared to using single classifiers. Note that the resulting model is a form of an ensemble of ensembles. Three key directions can be considered for future work. The first direction focuses on enhancing the scalability of the suggested hybrid model to address big datasets in terms of the number of instances and classes. The second direction concentrates on investigating more application domains to evaluate the generalization of the hybrid model. The third direction focuses on reducing the model complexity by exploring the integration of pruning techniques.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Esra'a Alshdaifat	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	
Fairouz Hussein		✓		✓	✓				✓	✓	✓			
Ala'a Al-shdaifat	✓		✓	✓				✓		✓	✓			
Malak Al-Hassan		✓				✓	✓	✓		✓				
Enshirah Altarawneh		✓				✓	✓	✓		✓				

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in the University of California Irvine (UCI) Machine Learning Repository, at https://archive.ics.uci.edu/.

REFERENCES

[1] P. Kaur and A. Gosain, "Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise," in *ICT Based Innovations*, Springer, Singapore, 2018, pp. 23–30. doi: 10.1007/978-981-10-6602-3_3.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[3] Haibo He, Yang Bai, E. A. Garcia, and Shutao Li, "ADASYN: adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, IEEE, Jun. 2008, pp. 1322–1328. doi: 10.1109/IJCNN.2008.4633969.

[4] Y. Yang, H. A. Khorshidi, and U. Aickelin, "A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems," *Frontiers in Digital Health*, vol. 6, Jul. 2024, doi: 10.3389/fdgth.2024.1430245.

[5] A. Özdemir, K. Polat, and A. Alhudhaif, "Classification of imbalanced hyperspectral images using smote-based deep learning methods," *Expert Systems with Applications*, vol. 178, p. 114986, Sep. 2021, doi: 10.1016/j.eswa.2021.114986.

[6] M. Waqar, H. Dawood, H. Dawood, N. Majeed, A. Banjar, and R. Alharbey, "An efficient smote-based deep learning model for heart attack prediction," *Scientific Programming*, vol. 2021, pp. 1–12, Mar. 2021, doi: 10.1155/2021/6621622.

[7] Z. Zheng, Y. Cai, and Y. Li, "Oversampling method for imbalanced classification," *Computing and Informatics*, vol. 34, no. 5, pp. 1017–1037, 2016.

[8] M. A. U. H. Tahir, S. Asghar, A. Manzoor, and M. A. Noor, "A classification model for class imbalance dataset using genetic programming," *IEEE Access*, vol. 7, pp. 71013–71037, 2019, doi: 10.1109/ACCESS.2019.2915611.

[9] M. A. Arefeen, S. T. Nimi, and M. S. Rahman, "Neural network-based undersampling techniques," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 2, pp. 1111–1120, Feb. 2022, doi: 10.1109/TSMC.2020.3016283.

[10] M. Koziarski, "Radial-based undersampling algorithm for classification of breast cancer histopathological images affected by data imbalance," in *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, IEEE, Oct. 2019, pp. 1–5. doi: 10.1109/CISP-BMEI48845.2019.8966010.

[11] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, "Boosting methods for multi-class imbalanced data classification: an experimental review," *Journal of Big Data*, vol. 7, no. 1, p. 70, Dec. 2020, doi: 10.1186/s40537-020-00349-y.

[12] R. Panthong, "Combining smote and ova with deep learning and ensemble classifiers for multiclass imbalanced," *Journal of Computer Science*, vol. 18, no. 8, pp. 732–742, Aug. 2022, doi: 10.3844/jcssp.2022.732.742.

[13] H. Guan, "A novel imbalanced classification method based on decision tree and bagging," *International Journal of Performability Engineering*, vol. 14, no. 6, pp. 1140–1148, 2018, doi: 10.23940/ijpe.18.06.p5.11401148.

[14] S. Kumar, P. Kaur, and A. Gosain, "A comprehensive survey on ensemble methods," in *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, IEEE, Apr. 2022, pp. 1–7. doi: 10.1109/I2CT54291.2022.9825269.

[15] R. K. Sevakula and N. K. Verma, "Balanced binary search tree multiclass decomposition method with possible non-outliers," *SN Applied Sciences*, vol. 2, no. 6, p. 1130, Jun. 2020, doi: 10.1007/s42452-020-2853-6.

[16] N. Junsomboon and T. Phienthrakul, "Combining over-sampling and under-sampling techniques for imbalance dataset," in *Proceedings of the 9th International Conference on Machine Learning and Computing*, New York, NY, USA: ACM, Feb. 2017, pp. 243–247. doi: 10.1145/3055635.3056643.

[17] J. Sun, J. Lang, H. Fujita, and H. Li, "Imbalanced enterprise credit evaluation with dte-sbd: decision tree ensemble based on smote and bagging with differentiated sampling rates," *Information Sciences*, vol. 425, pp. 76–91, Jan. 2018, doi: 10.1016/j.ins.2017.10.017.

[18] S. Shi, J. Li, D. Zhu, F. Yang, and Y. Xu, "A hybrid imbalanced classification model based on data density," *Information Sciences*, vol. 624, pp. 50–67, May 2023, doi: 10.1016/j.ins.2022.12.046.

[19] K. Theephoowiang and A. Hanskunatai, "A partition-based hybrid algorithm for effective imbalanced classification," *Data*, vol. 10, no. 4, p. 54, Apr. 2025, doi: 10.3390/data10040054.

[20] A. Shan and Y.-C. Chung, "A hybrid model based on samples difficulty for imbalanced data classification," in *Artificial Neural Networks and Machine Learning – ICANN 2023*, Springer, Cham, 2023, pp. 26–37. doi: 10.1007/978-3-031-44207-0_3.

[21] A. Salehi and M. Khedmati, "Hybrid clustering strategies for effective oversampling and undersampling in multiclass classification," *Scientific Reports*, vol. 15, no. 1, p. 3460, Jan. 2025, doi: 10.1038/s41598-024-84786-2.

[22] S. Wang and X. Yao, "Diversity analysis on imbalanced data sets by using ensemble models," in *2009 IEEE Symposium on Computational Intelligence and Data Mining*, IEEE, Mar. 2009, pp. 324–331. doi: 10.1109/CIDM.2009.4938667.

[23] H. Hartono, M. K. Zuhanda, R. Syah, S. Rahman, and E. Ongko, "A hybrid gdhs and gbdt approach for handling multi-class imbalanced data classification," *International Journal of Engineering, Science and Information Technology*, vol. 5, no. 3, pp. 51–57, May 2025, doi: 10.52088/ijesty.v5i3.894.

[24] J. Sun and J. Zhu, "Multi-class imbalanced corporate bond default risk prediction based on the ovo-smote-adaboost ensemble model," in *Proceedings of CECNet 2021*, IOS Press, Dec. 2021, pp. 42–53. doi: 10.3233/FAIA210388.

[25] X. Gao *et al.*, "A multiclass classification using one-versus-all approach with the differential partition sampling ensemble," *Engineering Applications of Artificial Intelligence*, vol. 97, p. 104034, Jan. 2021, doi: 10.1016/j.engappai.2020.104034.

[26] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques (the morgan kaufmann series in data management systems)*, 3rd ed. Morgan Kaufmann, 2011.

[27] M. Kelly, R. Longjohn, and K. Nottingham, "The uci machine learning repository." Accessed: Jan. 05, 2024. [Online]. Available: https://archive.ics.uci.edu/

[28] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86–92, Mar. 1940, doi: 10.1214/aoms/1177731944.
[29] P. B. Nemenyi, "Distribution-free multiple comparisons," Princeton University, 1963.

BIOGRAPHIES OF AUTHORS

Dr. Esra'a Alshdaifat earned her Ph.D. in Computer Science from the University of Liverpool, UK, in 2015, her M.Sc. in Computer Information Systems from Yarmouk University, Jordan, in 2008, and her B.Sc. in Computer Information Systems from the Hashemite University, Jordan, in 2006. Currently, she is an Associate Professor in the Department of Information Technology, Program of Data Science and AI at the Hashemite University, Zarqa, Jordan. With over 15 years of teaching experience. Her research interests include knowledge discovery in databases (KDD), data mining, machine learning, pattern recognition, natural language processing (NLP), and information retrieval. She can be contacted at email: esraa@hu.edu.jo.

Dr. Fairouz Hussein is a full-time lecturer at Crown Institute of Higher Education (CIHE), where she imparts her extensive knowledge and expertise to students. Additionally, she holds the position of Associate Professor at Hashemite University (HU), where she has dedicated more than 18 years to teaching and academic development. Throughout her career, she has supervised master's students, examined master and Ph.D. theses, and developed multiple courses. She earned her Ph.D. from the University of Technology Sydney (UTS), focusing her dissertation on action recognition and video summarization by submodular inference, which significantly contributed to her field. Her research interests encompass machine learning, cybersecurity, computer vision, image processing, greedy algorithms, and multimedia. She can be contacted at email: fairouz.hussein@cihe.edu.au.

Ms. Ala'a Al-shdaifat received the B.Sc. degree in computer science from the Hashemite university, Jordan, the M.Sc. degree in Information System from the University of Jordan, Jordan. Her research interests include data mining, machine learning, pattern recognition, natural language processing (NLP), and information retrieval. She can be contacted at email: alaa_shdaifat@hu.edu.jo.

Dr. Malak Al-Hassan is an Associate Professor in the Department of Business Information Technology at The University of Jordan, where she has been teaching since 2015. She holds a Ph.D. in Computer Information Systems from the University of Technology, Sydney, with a specialization in Intelligent Systems and E-services. She also earned a master's degree from the Jordan University of Science and Technology and a bachelor's degree from Yarmouk University. She is an active researcher, with her work advancing the fields of intelligent systems and e-government services. Her key research areas include e-services and e-government, semantic-enhanced recommender systems, sentiment analysis, web intelligence analysis, and decision support systems. She can be contacted at email: m_alhassan@ju.edu.jo.

Dr. Enshirah Altarawneh is an Assistant Professor in the Department of Computer Engineering at Hashemite University. She holds a Ph.D. in Computer Engineering from the State University of New York at Binghamton, specializing in cybersecurity and digital forensics. Her research focuses on artificial intelligence, machine learning, cybersecurity, and digital forensics. She has served as Department Chair, overseeing curriculum development, student advisement, faculty management, and other responsibilities. She also served as graduate studies dean assistant and is currently the students' affairs dean assistant. Additionally, she is an active member of IEEE. She can be contacted at email: enshirah@hu.edu.jo.