



# Investigating Accuracy and Diversity in Heterogeneous Ensembles for Breast Cancer Classification

Bouchra El Ouassif<sup>1</sup>, Ali Idri<sup>1,2</sup>(✉), and Mohamed Hosni<sup>1,3</sup>

<sup>1</sup> Software Project Management Research Team, ENSIAS, Mohammed V University,  
Rabat, Morocco

{bouchra.elouassif, ali.idri}@um5.ac.ma

<sup>2</sup> MSDA, Mohammed VI Polytechnic University, Ben Guerir, Morocco

<sup>3</sup> MOSI, L2M3S, ENSAM-Meknes, Moulay Ismail University, Meknes, Morocco

**Abstract.** Breast Cancer (BC) is one of the most common forms of cancer among women. Detecting and accurately diagnosing breast cancer at an early phase increase the chances of women's survival. For this purpose, various single classification techniques have been investigated to diagnosis BC. Nevertheless, none of them proved to be accurate in all circumstances. Recently, a promising approach called ensemble classifiers have been widely used to assist physicians accurately diagnose BC. Ensemble classifiers consist on combining a set of single classifiers by means of an aggregation layer. The literature in general shows that ensemble techniques outperformed single ones when ensemble members are accurate (i.e. have the lowest percentage error) and diverse (i.e. the single classifiers make uncorrelated errors on new instances). Hence, selecting ensemble members is often a crucial task since it can lead to the opposite: single techniques outperformed their ensemble. This paper evaluates and compares ensemble members' selection based on accuracy and diversity with ensemble members' selection based on accuracy only. A comparison with ensembles without member selection was also performed. Ensemble performance was assessed in terms of accuracy, F1-score. Q statistics diversity measure was used to calculate the classifiers diversity. The experiments were carried out on three well-known BC datasets available from online repositories. Seven single classifiers were used in our experiments. Skott Knott test and Borda Count voting system were used to assess the significance of the performance differences and rank ensembles according to theirs performances. The findings of this study suggest that: (1) Investigating both accuracy and diversity to select ensemble members often led to better performance, and (2) In general, selecting ensemble members using accuracy and/or diversity led to better ensemble performance than constructing ensembles without members' selection.

**Keywords:** Breast cancer · Classification · Combining classifiers · Heterogeneous ensemble · Diversity measures · Voting

## 1 Introduction

Breast Cancer is one of the most prominent diseases prevalent in females. It is the most common invasive cancer. In 2020, 2.3 million cases of BC were reported, which makes this cancer the most common cancer [1]. BC tumors occur when certain breast cells grow out of control, leading to the lining of the breast ducts. In addition, tumors can be classified into benign and malignant [2]. When cells are not cancerous, the tumor is benign. It would not invade nearby tissues or spread to other areas of the body (metastasize). When removed, benign tumors usually do not grow back. Malignant tumors, however, can grow and spread to other parts of the body.

Early detection of BC is the key to increase the survival rate, and the cancer is more likely to respond to effective treatment and can result in a greater probability of surviving, less morbidity, and less expensive treatment. Different techniques were investigated to diagnosis BC such as ultrasound, thermography, mammography and biopsy [3]. Mammography is possibly the most intensely used medical procedure by the physicians. However, if a mammogram looks suspicious, then a biopsy is required to decide whether an abnormality is in fact a breast cancer [4]. Furthermore, if the detection accuracy of the tumor is at a low level (lower than 70%), this prevents the doctor from reporting the final result of the diagnosis. Thus, this results in a waste of time and can cause mental discomfort for the patient [5]. Accordingly, the patient faces additional tests which can be costly and demanding.

In order to provide a quick and accurate diagnosis for BC [6], various classification techniques have been investigated in the literature such as Neural Networks (ANNs), Support Vector Machines (SVMs) and K Nearest Neighbor (KNN) [7, 8]. Nevertheless, single techniques are not always the most appropriate techniques to use, since they does not achieve better performance under all situations. Actually, the performance of single techniques relies on the characteristics of the dataset [5, 8, 9]. Furthermore, each single classification technique has advantages and limitations regarding the classification tasks. In order to address this challenge, a powerful approach called ensemble classifiers has been widely investigated. They consist of combining a set of individual classifiers by means of an aggregation layer [8, 10].

One of the most important task in optimizing an ensemble learning system is to select a subset of the “best” classifiers (ensemble members) from the whole pool of classifiers, which can drive an ensemble to outperforming its members [8, 11, 12]. Otherwise, the performance of an ensemble can be worse than all or most of its members [13]. In the literature, several previous studies were carried out to identifying optimal ways to combine classifiers [14–17]. However, the selection of the ensemble members is also a crucial challenge to deal with in order to improve the ensemble performance [18]. For instance, in [15] Aytu et al., proposed a hybrid ensemble approach that employs randomized search and clustering scheme to produce an ensemble. They trained a multitude of single classifiers with different parameters, then a group of diverse classifiers is created. Classifier clusters are then created using the classification performance of single classifiers. Thereafter, two single classifiers from each cluster are selected as candidate members based on their pairwise diversity to generate the ensemble. Caruana et al. [16] proposed an ensemble selection scheme from a library of thousands of classification algorithms. In this scheme, many machine learning algorithms and parameter settings

are used to build a model library. Then, a selection strategy, such as the forward step-wise selection, was used to select members that maximize the ensemble performance. In [17], Aksela proposed a method to select ensemble members based on several selection criteria such as correlation between errors, Q statistics and weighted count of errors. A measure focused on penalizing classifiers making the same error, the exponential error count approach, was identified to generate the best selections.

Performance of an ensemble learning can be influenced by many criteria including accuracy of single classifiers, number of base classifiers, combination rule, data sampling technique, and diversity of members [19]. While diversity was in general considered as the relevant criterion impacting the performance of an ensemble [20–23], other studies confirmed the opposite [24–27]. Diversity alone is a poor predictor of the ensemble accuracy” [28]. In [29], Krogh and Vedelsby underlined that members could improve the performance when they are accurate and diverse.

In the literature, several existing selection methods of ensemble members are essentially investigating one criterion: accuracy or diversity [18, 30–33]. This paper proposes a method for selecting heterogenous ensemble members for breast cancer classification, which uses both accuracy and diversity as selection criteria. Accuracy is measured in terms of recall precision, and accuracy metrics, while diversity is evaluated by means of the Q statistic diversity measure, which is one of the most popular due to its simplicity and understandability compared to other diversity measures [34].

Furthermore, we compare our proposed selection method with two existing strategies: (1) the selection of ensemble members based only on the criterion accuracy (i.e. we investigated the effect of the selection of the most accurate models from a group of seven classifiers); and (2) the selection of all the single classifiers without using any criterion. The empirical evaluations were carried out using: (1) seven single classifiers: K nearest neighbor (KNN), Multilayer Perceptron (MLP), Decision trees (DTs) and four variants of Support vector machines (SVMs) with four different kernels: Linear Kernel (LK), Normalized Polynomial Kernel (NP), Radial Basis Function Kernel (RBF), and Pearson VII function based Universal Kernel (PUK); (2) a majority voting combination rule to combine the outputs of the ensemble members. (3) three well-known available BC datasets from online repositories; (4) three performance metrics, namely accuracy, recall and precision to evaluate the constructed ensembles; and (5) the statistical test Scott-Knott and the Borda Count voting system to perform the significance tests and rank the best classifiers respectively.

The contributions of this paper are: (1) analyzing the impact of investigating both accuracy and diversity for ensemble members’ selection in breast cancer classification; and (2) comparing the proposed members’ selection method with two existing selection strategies.

The rest of this paper is structured as follows: Sect. 2 briefly presents the single techniques used, the ensemble concept and the existing measures of diversity. Section 3 presents an overview of related work investigating diversity in members’ selection. Section 4 describes the experimental design pursued in this study. The empirical findings are presented and discussed in Sect. 5. The threats to validity are given in Sect. 6. Conclusions and future works are summarized in Sect. 7.

## 2 Background

This section gives a summary of the single classification techniques used, the concepts of ensemble, and the measures of diversity to select ensemble members in classification.

### 2.1 Single Techniques

**KNN:** is a popular machine learning algorithm known for its simple implementation and robustness [35]. It is a non-parametric method first created in 1951 by Evelyn Fix and Joseph Hodges, and later updated by Thomas Cover. It used to solve both classification and regression problems. KNN stores all available instances and classifies new instances based on a similarity measure. To measure the similarity between its nearest neighbors, KNN uses in general the Euclidian distance.

**SVMs:** are powerful classification algorithms, used to solve problems of classification as well as regression. SVM was developed in the 1990s by Vladimir Vapnik [36]. It is used to classify a new unknown instance into one of the predefined classes. SVM has the ability to model complex nonlinear relationships by choosing an appropriate kernel function [37]. In fact, the Kernel function transforms the training samples so that a non-linear decision boundary is transformed to a linear equation in a higher number of dimensions [12, 38]. In this study, four variants of the SVM classifier were used. The four SVMs variants used four different kernels: Linear Kernel (LK), Radial Basis Function Kernel (RBF, Pearson VII function based Universal Kernel (PUK) and Normalized Polynomial Kernel (NP).

**MLP Neural Networks:** are the most frequently used feedforward neural networks due to their fast operation, ease of implementation, and smaller training set requirements, [39, 40]. They are used for both classification and regression problems [41, 42]. Their architecture consists of three types of layers: the input layer, output layer and hidden layer. The nodes present in each layer are connected to the next layer. That is the principle of feed-forward neural network; the movement information is allowed only in a forward direction. The neurons of each layer are connected to the neurons of the subsequent layer by means of weights and output signals which are a function of the sum of the inputs to neurons modified by an activation function. Generally, the neurons of the hidden layer use a nonlinear activation function, while a linear activation function is usually used for the output neurons.

**DTs:** are the most frequently used classification techniques, easy to use and to interpret. They can be used for both classification and regression problems [43]. DT is a tree-structured model in which internal nodes represent dataset attributes, branches represent decision rules, and each leaf node represents a class label. Depending on the task addressed, the class label could be categorical or continuous. The classification rules are described by the paths from root to leaf. In this study, the C4.5 algorithm was investigated [44].

## 2.2 Ensemble Classifiers

An ensemble classifiers are a powerful machine learning technique that create multiple models and then combine them by means of an aggregation rule in order to produce one optimal predictive model [8, 12]. They can be grouped into two types: Homogeneous or Heterogeneous [8, 10, 12, 45]. The Homogeneous method refers to an ensemble that combines one based learning algorithm with at least two different variants, or an ensemble that combines one base learning algorithm with one meta ensemble such as Boosting [46]. While the Heterogeneous method refers to an ensemble that combines members having different base learning algorithms. The current research is based on heterogeneous ensembles, and it adopts the majority voting combination rule to combine the decision of the individual classifiers that comprise the ensemble. Note that, the majority vote rule is the most popular and frequently used method in the literature of ensembles [47].

## 2.3 Measures of Diversity in Ensemble Based Classification

It is well known that the performance of an ensemble learning is impacted by diversity of its members, i.e., the degree of disagreement within the members of an ensemble [18, 25]. Diversity is loosely described as “making errors on different examples” [48, 49]. Thus, diversity has been acknowledged as a very relevant characteristic in classifiers combination. Kuncheva [28, 50] provided an analysis of ten diversity measures and classified them into two groups: Pairwise and Non-pairwise measures. Pairwise measures calculate diversity values between two base classifiers of an ensemble. The overall diversity of an ensemble can be estimated by averaging the pairwise diversity values of pairs using Q-statistic [51], double-default measure [52], and disagreement measure [50]. Non-pairwise measures, on the other hand, are used to estimate diversity among all base classifiers by accounting for all potential disagreements between them using entropy [53], generalized diversity [54], and measure of difficulty [50]. In this study, the Q statistic diversity measure was used, it is preferred over other diversity measures because of its simplicity and understandability [34].

Q statistic measure is based on Yule’s Q statistic used to assess the similarity of two classifiers’ outputs [50]. For two classifiers  $L_i$  and  $L_k$ , Q-statistic value is defined by Eq. 1.

$$Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \quad (1)$$

where  $N_{ab}$  is the number of training instances for which  $L_i$  gives result ‘a’ and  $L_k$  gives result ‘b’ (It is supposed that the result here is equal to 1 if an instance is classified correctly and 0 if it is misclassified). The expected value of Q for statistically independent classifiers is 0. The value of Q ranges between  $-1$  and  $1$ . Classifiers that appear to correctly classify the same instances will have positive Q values, while those that make errors on different instances will have negative Q values [21, 50].

### 3 Related Work

This section presents an overview of some related work investigating diversity in ensemble techniques.

Banfield et al. [55] proposed an algorithm termed the percentage correct diversity measure (PCDM) to construct decision trees ensemble. The proposed algorithm seeks to find the test samples for which the percentage between 10 and 90 of the single classifiers are correct. These test samples are removed from determining the ensemble diversity. The proposed technique was evaluated in twelve datasets from UCI repository using ten fold cross validation method. The empirical results suggest the effectiveness of the proposed technique.

Kadkhodaei et al. [56] proposed an entropy based approach to determine the best combination of classifiers from a pool of ten different single techniques. The evaluation of the proposed heterogeneous ensemble was evaluated on three datasets from the UCI repository. The empirical results stated that the proposed technique generates an accurate ensemble and that the time required to build it is less than the one required bagging and boosting ensemble techniques.

Nascimento et al. [57] presented a new approach for automatic selection of both base classifiers and features. The proposed approach was based on evolutionary approach composed of two genetic algorithm instances. Two proposed diversity measures were investigated in order to analyze the performance of the proposed framework. The empirical evaluations were performed using ten different classification algorithms using the bagging architecture. Five datasets from UCI repository were selected for the evaluations. The results suggested that the proposed technique was effective to generate accurate ensemble. The authors recommended to take into account other factors than diversity such as accuracy and complexity when constructing ensembles.

Lysiak et al. [58] proposed a novel approach for dynamic ensemble selection (DES) based on probabilistic measures of competence and diversity between member classifiers. The two types of ensembles were constructed: the homogeneous ensemble consisted of 20 pruned decision tree classifiers and the heterogeneous ensemble consisted of nine different classifiers. Seven public datasets were used to assess the effectiveness of the proposed approach. The Results indicated that the proposed method can eliminate weak classifiers and keep the ensemble maximally diverse. Further, the proposed DES led to better classification accuracy of the constructed ensembles compared to those generated by the DES system using only the competence measure.

## 4 Experimental Design

This section explains the experimental design investigated to conduct all of the empirical evaluations, including the performance metrics used, Scott–Knott (SK) test, Borda Count voting system, datasets descriptions, ensemble selection process and the abbreviations used.

### 4.1 Performance Metrics

The following performance metrics are used to assess the performance of single and ensemble techniques:

Accuracy, Recall and Precision defined by Eqs. 2, 3 and 4 respectively [10].

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{Precision(Prec)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

where FP stands to False Positive, FN stands to False Negative TP to True Positive and TN to True Negative.

### 4.2 Scott-Knott Test

The Scott-Knott (SK) test is a hierarchical clustering algorithm developed by Scott and Knott (1974), is an efficient method to conduct procedures of multiple comparisons without ambiguity [59]. Compared to other statistical tests such as the Tukey test, Student–Newman–Keuls (SNK) test and t-test, the SK test is a commonly used method [60–63], it has the ability to group techniques into non-ambiguous groups [64, 65]. In this study, the SK test was used to cluster the single and ensemble techniques based on their error rates (Error rate = 1-Accuracy) and to check the significant difference between them. The ten folds cross-validation approach was used in all the experiments presented in this study.

### 4.3 Borda Count Voting System

The Borda count [66] is a form of single-winner election in which voters rate candidates in order of choice. The Borda count determines the winner of an election by allocating points to each candidate based on the voter’s rating. After that, the system aggregates the score of each candidate based on the received points. The candidate who receives the highest score is the winner. To illustrate this process, the example in Table 1 shows the steps of Borda count to choose the winner among four candidates (Cd<sub>1</sub>, Cd<sub>2</sub>, Cd<sub>3</sub> and Cd<sub>4</sub>) who were voted on by four voters (Vot<sub>1</sub>, Vot<sub>2</sub>, Vot<sub>3</sub> and Vot<sub>4</sub>). Each voter Vot<sub>i</sub> assigns a candidate to one of the positions i (1, 2, 3 or 4). As a result, we compute the

**Table 1.** Borda count voting system for four voters which rank four candidates according to their preferences.

Voters	VOT <sub>1</sub>	VOT <sub>2</sub>	VOT <sub>3</sub>	VOT <sub>4</sub>	POS I	Score
Cd <sub>1</sub>	1	1	2	4	2, 1,0,1	4 × 2 + 3 × 1 + 2 × 0 + 1 × 1 = 12
Cd <sub>2</sub>	3	2	3	2	0,2,2,0	4 × 0 + 3 × 1 + 2 × 2 + 1 × 0 = 7
Cd <sub>3</sub>	1	2	1	4	2,1,0,1	4 × 2 + 3 × 1 + 1 × 2 + 1 × 1 = 14
Cd <sub>4</sub>	4	2	3	1	1,1,1,1	4 × 1 + 3 × 1 + 1 × 2 + 1 × 1 = 10

vector position PI  $(n_1, n_2, n_3, n_4)$  of each candidate, where  $n_i$  is the number of times the candidate has been ranked in position  $i$ .  $M_i$  points are assigned to each position  $i$  ( $M_i = \# \text{ of candidates} - i + 1$ ). Finally, the score of each candidate is equal to  $n_i * M_i$ .

#### 4.4 Datasets Fescription

In order to evaluate the performance of the proposed techniques, three datasets obtained from the online UCI repository were investigated in this study. These datasets were the most widely used by researchers in the literature [7]. A short description of each of these datasets is reported in Table 2. Note that two of the datasets contain missing values. We simply removed them since their number was very small. Furthermore, the WPBC and Wisconsin datasets are unbalanced. The Synthetic Minority Over-sampling Technique (SMOTE) [67] was used to address this problem.

**Table 2.** Datasets description

Dataset	#.Attributes	Missing values?	Examples
WDBC	32	NO	569
Wisconsin	11	Yes(16)	699
WPBC	34	Yes (4)	198

#### 4.5 Ensemble Selection Process

The first concern of constructing an effective ensemble classifier is to ensure that all individual classifiers are accurate [68]. Then we can improve the ensemble performance by rejecting weak classifiers and combining accurate members only. Toward this aim, we select ensemble members based on accuracy, referred to us Selection by Accuracy (SbA). The process of SbA is as follow:

1. Assess the performances of the N (7 in this study) single classifiers based on the three metrics: Accuracy, Precision and Recall.
2. Performing the statistical test SK based on the accuracy in order to cluster the classifiers evaluated in Step 1 into non-overlapping clusters. Each cluster includes one or more classifiers with comparable predictive abilities. The best cluster is the one with the lowest error rate value. (Error rate = 1-Accuracy); therefore, the classifiers belonging to this cluster are chosen for the next step.
3. Building an heterogeneous ensemble by combing the base classifiers belonging to the best cluster (In case of the best cluster only contains one classifier, we combine the classifiers of the two first best clusters)
- 4 Evaluate the performance of the heterogeneous ensemble constructed in Step 3 according to the three criteria: accuracy, recall and precision.



Since diversity of the ensemble members is relevant to improve the accuracy of an ensemble, we construct ensembles based on both accuracy and diversity, referred to us Selection by Accuracy and Diversity (SbAD). The process of SbAD is as follows.

1. Carry out the Step 1, 2 and 3 of the SbA selection process.
2. Calculate the diversity between the heterogeneous ensemble classifier constructed in Step 1 (i.e. based on accuracy criterion) and each of the remaining classifiers (i.e. classifiers that were out of the best cluster)
3. Select the single classifier with the highest diversity and include it into the SbA heterogeneous ensemble.
4. Assess the performance of the heterogeneous ensemble constructed in Step 3.
5. Repeat Steps 3 and 4 until the heterogeneous ensemble size reaches the number  $N-1$  members.

We also combine all the seven single classifiers (DT, MLP, KNN, S-PUK, SVM-NP, SVM-RBF and SVM-LK) in one ensemble in order to compare its performance with the other constructed ensembles using SbA and SbAD. We referred to this ensemble: No Selection classifier (NSc). Figure 1 presents the experimental process we followed.

#### 4.6 Abbreviations Used

The following abbreviation rules were used to simplify the names of ensembles

*E-SingleTechnique1 SingleTechnique2*  
*E- SingleTechnique1 SingleTechnique2SingleTechnique3.*  
 .  
 .  
*E- SingleTechnique1 SingleTechnique2... SingleTechniqueN*

It is worth noting that for ensemble techniques, we shorten the names of single classifiers as well:

KNN for K, D for DTs, M for MLP, S for SVM, SVM-PUK for P, SVM-RBF for R, S-LK for L and SVM-NP for NP.

For example, EDKLM refers to the ensemble constructed by the fusion of the four single techniques, DT, KNN, SVM-LK and MLP.

## 5 Empirical Results

This section discusses the empirical evaluation results of the 7 individual classifiers, the SbA and SbAD ensembles as well as the Nsc ensemble. The R software was used for statistical tests and the Waikato Environment for Knowledge Analysis (WEKA 3.9) was investigated to conduct the empirical evaluations [69].

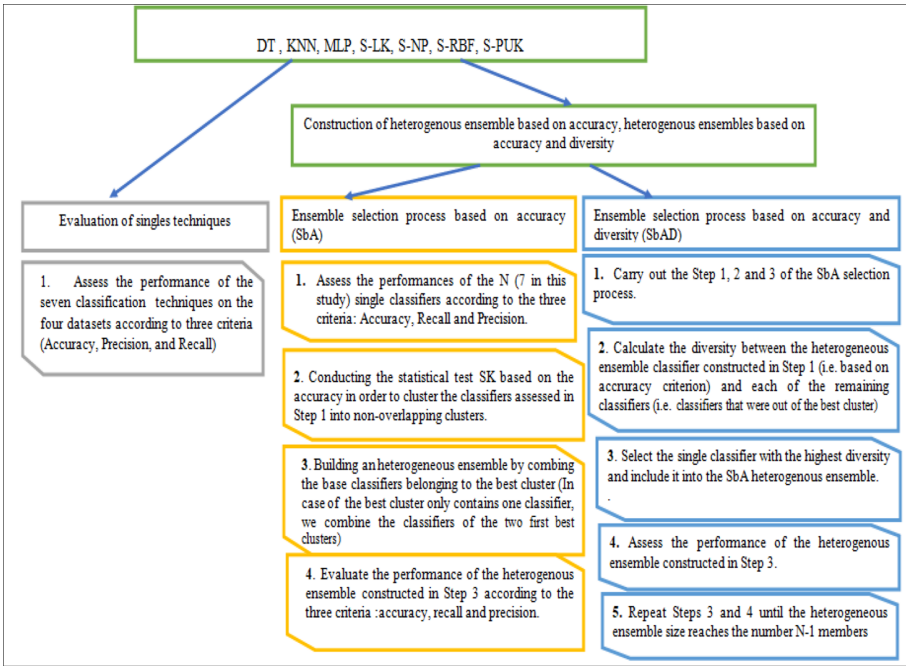


Fig. 1. Experimental process.

5.1 Individual Classifiers Evaluation

This section presents the evaluation results of the 7 classification techniques over the three datasets based on three criteria (Accuracy, Recall, and Precision). Table 3 depicts the performance values of the 7 single techniques. As shown in Table 3, we can note that in the WDBC dataset, S-LK displayed the best performance. S-PUK exhibited the best performance in the WPBC dataset. As for Wisconsin, DT was the best.

Based on the results of Table 3, we notice that there is no individual best classifier in overall datasets, and that the rankings of the same individual classifier vary depending on the dataset. This is due to the fact that the performance of single techniques depends on the characteristics of datasets (number of instances, dataset dimensionality, number of classes, etc.) [8].

Table 3. Performance results.

Classifier	WDBC			Wisconsin			WPBC		
	Acc	Prec	Recall	Acc	Prec	Recall	Acc	Prec	Recall
S-LK	97.89	97.9	97.9	74.66	74.8	74.7	74.66	74.8	74.7
S-NP	93.4	93.5	93.5	77.03	77	77.2	77.03	77	77.2

(continued)

**Table 3.** (continued)

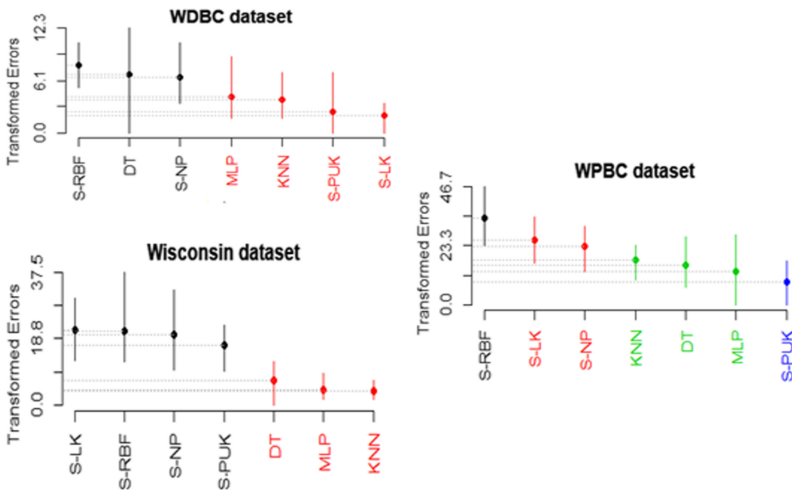
Classifier	WDBC			Wisconsin			WPBC		
	Acc	Prec	Recall	Acc	Prec	Recall	Acc	Prec	Recall
S-RBF	92.09	92.1	92.8	65.88	65.9	66.5	65.88	65.9	66.5
S-PUK	97.54	97.5	97.5	90.88	90.9	90.9	<b>90.88</b>	<b>90.9</b>	<b>90.9</b>
DT	93.15	93.2	93.1	<b>96.05</b>	<b>96.1</b>	<b>96</b>	84.12	84.1	84.1
MLP	95.78	95.8	95.8	95.75	95.8	95.8	86.82	88.1	86.8
KNN	95.78	96.1	96.1	95.75	95.8	95.8	82.43	85	82.4

**5.2 Ensembles Evaluation**

Figure 2 depicts the results of the SK test carried out based on error rate overall datasets. We observed that the SK test identified 2, 4, and 2 clusters in the WDBC, WPBC and Wisconsin datasets respectively. Therefore, we constructed: (1) one SbA heterogeneous ensemble in WDBC dataset whose members were S-PUK, S-LK, KNN and DT; (2) one SbA heterogeneous ensemble in Wisconsin dataset whose members were DT, MLP and KNN; and (3) one SbA heterogeneous ensemble in WPBC dataset whose members were S-PUK, MLP, DT and KNN (since the best cluster of WPBC dataset contains only one technique we include also the techniques of the second best cluster).

For the SbAD ensembles, for each dataset we constructed ensembles based on the SbAD selection process described above (see Sect. 4.5):

- For Wisconsin, we constructed 3 SbAD ensembles: EDKNM, EDKLM and EDKLNMP



**Fig. 2.** SK test of single techniques over the all datasets

- For WDBC, we constructed 2 SbAD ensembles: EDLKMP and EDLKMNP
- For WPBC, we constructed 2 SbAD ensembles: EDKMPR and EDKMPRN

**WDBC Dataset:** Table 4 shows the performance metric values of the SbA and SbAD ensembles as well as the NSc ensemble for the WDBC dataset. We observe that the SbAD ensemble EDLKMP and the SbA ensemble come first with **97.72%**, **97%** and **97%** for accuracy, recall and precision respectively; and the SbAD ensemble EDLKMNP comes second with 97.54%, 97.5% and 97.6% for accuracy, recall and precision respectively. Note that the the SbA and SbAD ensembles outperformed the NSc ensemble.

**Table 4.** Performance results: WDBC dataset.

Method	Ensemble	Acc	Prec	Recall
SbDA	EDLKMP	<b>97.72</b>	<b>97</b>	97
	EDLKMNP	97.54	97.5	97.6
SbA	ELKMP	<b>97.72</b>	<b>97</b>	97
NSc		97.12	97.2	97.2

**Wisconsin Dataset:** Table 5 reports the performance metric values of the SbA and SbAD ensembles as well as the NSc ensemble for the Wisconsin dataset. In terms of the three performance metrics, accuracy, precision, and recall, we notice that the SbAD ensemble EDKLNMP marginally outperformed the others. It provides an accuracy, precision and recall values of 97.07%, 97.1% and 97.1% respectively.

**Table 5.** Performance results: Wisconsin dataset.

Method	Ensemble	Acc	Prec	Recall
SBAD	EDKNM	96.49	96.5	96.5
	EDKLNM	96.92	97	96.9
	EDKLNMP	<b>97.07</b>	<b>97.1</b>	<b>97.1</b>
SBA	EDKM	96.63	96.7	96.6
NSc		96.78	96.8	96.8

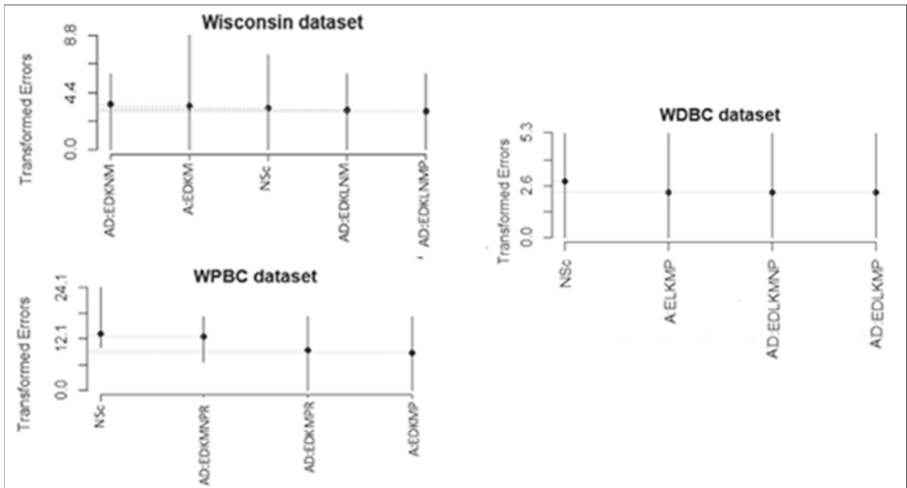
**WPBC dataset:** Table 6 depicts the performance metrics values of the SbA and SbAD ensembles as well as the NSc ensemble for the WPBC dataset. We observe that the ensemble SbA ensemble EDKMP outperformed all the other ensembles; it achieved an accuracy, precision and recall of 91.22%, 91.5% and 91.2% respectively. As it can be seen from Table 6, the SbA and SbAD ensembles in general outperformed the NSc ensemble.

**Table 6.** Performance results: WPBC dataset.

Method	Ensemble	Acc	Prec	Recall
SBAD	EDKMPR	90.54	90.8	90.5
	EDKMPRN	87.5	87.6	87.5
SBA	EDKMP	<b>91.22</b>	<b>91.5</b>	<b>91.2</b>
NSc		86.82	87.1	86.8

### 5.3 Comparing SbA, SbAD and Nsc Ensembles

To check the significant difference between the performances of SbA, SbAD and Nsc ensembles, the SK test was carried out based on error rate values to check whether there was a notable difference between the ensemble performances. Figure 3 displays the results of the SK test on the built ensembles for each dataset. As it can be observed in Fig. 3, in all datasets, only one cluster was identified by the SK test. This means that SbA, SbAD and Nsc ensembles show the same predictive capabilities in terms of accuracy in all datasets.

**Fig. 3.** Accuracy based SK test results of SbA, SBA and Nsc ensembles.

To deeply compare the predictive capabilities of SbA, SbAD and Nsc ensembles, we used the Borda Count voting system to rank them based on the three performance metrics: accuracy, precision, and recall. Table 7 displays the ranking supplied by the Borda count voting system for each dataset. We note the following:

- Except in the WPBC dataset, where the SbA ensemble EDKMP was ranked first, SbAD ensemble techniques were in general ranked at the first position in two datasets (EDLKMP in WDBC and EDKLNMP in Wisconsin).

**Table 7.** Borda Count ranks of the best SK cluster techniques

Rank	WDBC dataset	Wisconsin dataset	WPBC dataset
1	SbAD:EDLKMP <sup>a</sup>	SbAD:EDKLNMP	SbA:EDKMP
2	SbA:ELKMP <sup>a</sup>	SbAD:EDKLNMP	SbAD:EDKMPPR
3	SbAD:EDLKMNPP	NSc	SbAD:EDKMPPRN
4	NSc	SbA:EDKM	NSc
5		SbAD:EDKNM	

<sup>a</sup>The letter denotes the same rank

- In general, the SbAD and SbA ensembles outperformed the NSc ensemble overall datasets.
- The best SbAD ensemble in each dataset includes 5 to 6 single techniques (EDLKMP in WDBC, EDKLNMP in Wisconsin and EDKMPPR in WPBC).
- The members DT, KNN, MLP, and S-PUK were present in all the best SbAD/SbA ensembles (EDLKMP in WDBC, EDKLNMP in Wisconsin and EDKMP in WPBC).

## 6 Threats to Validity

This section discusses threats to this study's validity, with regard to internal, external and construct validity.

**Internal Validity:** This study used a 10-fold cross validation evaluation method, which is commonly used in machine learning [10]. Another internal threat is the presence of missing values in two datasets used, which may affect the performance of a model constructed using these datasets and increase the likelihood of drawing incorrect conclusions. In this research, instead of using imputation techniques like mean imputation or expectation-maximization [70] to impute missing values, we deleted the existing missing values because their number was small, but if there are a lot of missing values, we recommend the use of imputation methods [71].

Furthermore, the majority voting was used in this study to produce the final output of the proposed ensembles. Note that, the majority voting rule is the most popular and frequently used method in the literature of ensembles [72, 73]. However, we are aware that other combination schemes such as weighted majority voting, probabilistic and weighted sum can yield different outcomes. Thus, future experiments might look into the effect of other combiners on the predictive capability of the ensembles.

**External Validity:** The aim of the external threat is to know whether the results of this research work can be generalized to other contexts [74]. Thus, it is crucial to figure out how broadly the findings of this study can be applied. To counteract this threat, we selected three datasets that differ in size and number of features. Moreover, this study investigated only one diversity measure, Q statistic, to calculate the classifiers' diversity. Investigating other diversity measures such as double-default measure [52], disagreement measure [50] and measure of difficulty [50] is recommended so as to generalize the findings of this study.

**Construct Validity:** The construct validity aims at answering the question about the measurement validity [11], or more specifically, the reliability/credibility of the measures selected to assess the performance of the techniques. We overcome this constraint by using three performance metrics (Accuracy, Precision and Recall) in order to obtain results from various perspectives. As a result, all three criteria yielded the same results. To affirm or refute the reported results, it is required to investigate other performance metrics.

## 7 Conclusion and Future Work

This study assessed and compared the impacts of using accuracy and diversity instead of accuracy alone for selecting ensemble members on the performance of ensembles in BC classification. Moreover, it evaluated and compared the performances of ensembles with/out selection. We used the Q statistic measure to evaluate the diversity of ensemble members. The majority voting combination rule was used to combine the members of an ensemble. The experiments were carried out on three well-known available BC datasets from online repositories. The SK test and Borda Count were investigated to assess the significance of performance differences and to rank the ensembles respectively.

The findings were:

- (1) Investigating both accuracy and diversity to select ensemble members can improve the performance of an ensemble. This confirms the findings of [75].
- (2) Selecting ensemble members using accuracy and/or diversity in general led to better ensemble performance than constructing ensembles without members' selection. This confirms the results of Zhou et al. [76].
- (3) Individual classifiers' performance is influenced by the characteristics of the dataset as the ranks of the same technique depended on the dataset.

Ongoing work focuses on investigating other diversity measures and other combination rules to construct better heterogeneous ensembles in BC diagnosis.

## References

1. breastCancer. <https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/>. Accessed 16 Jan 2019
2. Breast Cancer Facts - National Breast Cancer Foundation. <https://www.nationalbreastcancer.org/breast-cancer-facts>. Accessed 11 Dec 2020
3. Luo, S.T., Cheng, B.W.: Diagnosing breast masses in digital mammography using feature selection and ensemble methods. *J. Med. Syst.* **36**, 569–577 (2012). <https://doi.org/10.1007/s10916-010-9518-8>
4. Chhatwal, J., Alagoz, O., Burnside, E.S., Burnside, E.S.: Optimal Breast biopsy decision-making based on mammographic features and demographic factors. *Oper. Res.* **58**(6), 1577–1591 (2010). <https://doi.org/10.1287/opre.1100.0877>

5. Kaushik, D., Kaur, K.: Application of data mining for high accuracy prediction of breast tissue biopsy results. In: 2016 3rd Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications. DIPDMWC 2016, pp. 40–45 (2016). <https://doi.org/10.1109/DIPDMWC.2016.7529361>
6. Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence (2019)
7. Idri, A., Chlioui, I., El Ouassif, B.: A systematic map of data analytics in breast cancer. In: Proceedings of the Australasian Computer Science Week Multiconference (2018). <https://doi.org/10.1145/3167918.3167930>
8. Idri, A., Bouchra, E.O., Hosni, M., Abnane, I.: Assessing the impact of parameters tuning in ensemble based breast Cancer classification. *Heal. Technol.* **10**(5), 1239–1255 (2020). <https://doi.org/10.1007/s12553-020-00453-2>
9. El Ouassif, B., Idri, A., Hosni, M.: Homogeneous ensemble based support vector machine in breast cancer diagnosis (2021). <https://doi.org/10.5220/0010230403520360>
10. Hosni, M., Abnane, I., Idri, A., Carrillo de Gea, J.M., Fernández-Alemán, J.L.: Reviewing ensemble classification methods in breast cancer. *Comput. Methods Programs Biomed.* **177**, 89–112 (2019)
11. Hosni, M., Idri, A., Abran, A., Nassif, A.B.: On the value of parameter tuning in heterogeneous ensembles effort estimation. *Soft. Comput.* **22**(18), 5977–6010 (2017). <https://doi.org/10.1007/s00500-017-2945-4>
12. El Ouassif, B., Idri, A., Hosni, M.: Homogeneous ensemble based support vector machine in breast cancer diagnosis - BIOSTEC 2021. In: HEALTHINF 2021 - 14th International Conference on Health Informatics, Proceedings; Part of 13th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2021. SciTePress (2021)
13. Yang, L.: Classifiers selection for ensemble learning based on accuracy and diversity. *Procedia Eng.* **15**, 4266–4270 (2011). <https://doi.org/10.1016/j.proeng.2011.08.800>
14. Alexandropoulos, S.-A., Aridas, C.K., Kotsiantis, S.B., Vrahatis, M.N.: Stacking strong ensembles of classifiers. In: MacIntyre, J., Maglogiannis, I., Iliadis, L., Pimenidis, E. (eds.) AIAI 2019. IAICT, vol. 559, pp. 545–556. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-19823-7\\_46](https://doi.org/10.1007/978-3-030-19823-7_46)
15. Onan, A., Korukoğlu, S., Bulut, H.: A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. *Inf. Process. Manag.* **53**, 814–833 (2017). <https://doi.org/10.1016/j.ipm.2017.02.008>
16. Caruana, R., Crew, G., Ksikes, A.: Ensemble selection from libraries of models. In: The Proceedings of ICML 2004 (2004)
17. Aksela, M.: Comparison of classifier selection methods for improving committee performance. In: Windeatt, T., Roli, F. (eds.) MCS 2003. LNCS, vol. 2709, pp. 84–93. Springer, Heidelberg (2003). [https://doi.org/10.1007/3-540-44938-8\\_9](https://doi.org/10.1007/3-540-44938-8_9)
18. Butler, H.K., Friend, M.A., Bauer, K.W., Bihl, T.J.: The effectiveness of using diversity to select multiple classifier systems with varying classification thresholds. *J. Algorithms Comput. Technol.* **12**, 187–199 (2018). <https://doi.org/10.1177/1748301818761132>
19. Bian, S., Wang, W.: Investigation on Diversity in Homogeneous and Heterogeneous Ensembles (2006)
20. Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.* **51**, 181–207 (2003)
21. Wang, S., Yao, X.: Relationships between diversity of classification ensembles and single-class performance measures. *IEEE Trans. Knowl. Data Eng.* **25**, 206–219 (2013). <https://doi.org/10.1109/TKDE.2011.207>
22. Tsymbal, A., Pechenizkiy, M., Cunningham, P.: Diversity in search strategies for ensemble feature selection. *Inf. Fusion.* **6**, 83–98 (2005). <https://doi.org/10.1016/j.inffus.2004.04.003>



23. Windeatt, T.: Diversity measures for multiple classifier system analysis and design. *Inf. Fusion*, **6**, 21–36 (2005). <https://doi.org/10.1016/j.inffus.2004.04.002>
24. Schapire, R.E.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.* **51**, 181–207 (2003)
25. Duin, R.P.W., Tax, D.M.J.: Experiments with classifier combining rules. In: Kittler, J., Roli, F. (eds.) *MCS 2000*. LNCS, vol. 1857, pp. 16–29. Springer, Heidelberg (2000). [https://doi.org/10.1007/3-540-45014-9\\_2](https://doi.org/10.1007/3-540-45014-9_2)
26. Skurichina, M., Kuncheva, L.I., Duin, R.P.W.: Bagging and boosting for the nearest mean classifier: effects of sample size on diversity and accuracy. In: Roli, F., Kittler, J. (eds.) *MCS 2002*. LNCS, vol. 2364, pp. 62–71. Springer, Heidelberg (2002). [https://doi.org/10.1007/3-540-45428-4\\_6](https://doi.org/10.1007/3-540-45428-4_6)
27. Webb, G.I., Zheng, Z.: Multistrategy ensemble learning: reducing error by combining ensemble learning techniques. *IEEE Trans. Knowl. Data Eng.* **16**, 980–991 (2004). <https://doi.org/10.1109/TKDE.2004.29>
28. Kuncheva, L.I.: That elusive diversity in classifier ensembles. In: Perales, F.J., Campilho, A.J.C., de la Blanca, N.P., Sanfeliu, A. (eds.) *IbPRIA 2003*. LNCS, vol. 2652, pp. 1126–1138. Springer, Heidelberg (2003). [https://doi.org/10.1007/978-3-540-44871-6\\_130](https://doi.org/10.1007/978-3-540-44871-6_130)
29. Krogh, A., Vedelsby, J.: Neural network ensembles, cross validation, and active learning. In: *Proceedings of the 7th International Conference on Neural Information Processing Systems* (1994)
30. Kuncheva, L.I., Skurichina, M., Duin, R.P.W.: An experimental study on diversity for bagging and boosting with linear classifiers. *Inf. Fusion*, **3**, 245–258 (2002). [https://doi.org/10.1016/S1566-2535\(02\)00093-3](https://doi.org/10.1016/S1566-2535(02)00093-3)
31. Narasimhamurthy, A.: Evaluation of diversity measures for binary classifier ensembles. In: Oza, N.C., Polikar, R., Kittler, J., Roli, F. (eds.) *MCS 2005*. LNCS, vol. 3541, pp. 267–277. Springer, Heidelberg (2005). [https://doi.org/10.1007/11494683\\_27](https://doi.org/10.1007/11494683_27)
32. Azizi, N., Farah, N., Sellami, M., Ennaji, A.: Using diversity in classifier set selection for Arabic handwritten recognition. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. LNCS, **5997**, 235–244 (2010). [https://doi.org/10.1007/978-3-642-12127-2\\_24](https://doi.org/10.1007/978-3-642-12127-2_24)
33. Naldi, M.C., Carvalho, A.C.P.L.F., Campello, R.J.G.B.: Cluster ensemble selection based on relative validity indexes (2013). <https://doi.org/10.1007/s10618-012-0290-x>
34. Kotsiantis, S., Kanellopoulos, D., Pintelas, P.: Handling imbalanced datasets: a review. *GESTS Int'l Trans. Comput. Sci. Eng.* **30**, 25–36 (2012)
35. Zhang, S., Li, X., Zong, M., Zhu, X., Wang, R.: Efficient kNN classification with different numbers of nearest neighbors. *IEEE Trans. Neural Networks Learn. Syst.* **29**, 1774–1785 (2018). <https://doi.org/10.1109/TNNLS.2017.2673241>
36. Vapnik, V.: Principles of risk minimization for learning theory. In: *Advances in Neural Information Processing Systems* (1992)
37. Schölkopf, B., Alexander, J.S.: Support Vector Machines, Regularization, Optimization, and Beyond. In: *Learning with Kernels*, pp. 1–27 (2001)
38. Bhavsar, H., Ganatra, A.: Radial basis polynomial kernel (RBPK): a generalized kernel for support vector machine. *Int. J. Comput. Sci. Inf. Secur.* **14**, 1–20 (2016)
39. Kocyigit, Y., Alkan, A., Erol, H.: Classification of EEG recordings by using fast independent component analysis and artificial neural network. *J. Med. Syst.* **32**, 17–20 (2008). <https://doi.org/10.1007/s10916-007-9102-z>
40. Übeyli, E.D.: Combined neural network model employing wavelet coefficients for EEG signals classification. *Digit. Signal Process. A Rev. J.* **19**, 297–308 (2009). <https://doi.org/10.1016/j.dsp.2008.07.004>

41. Idri, A., Khoshgoftaar, T., Abran, A.: Can neural networks be easily interpreted in software cost estimation? In: 2002 IEEE World Congress on Computational Intelligence. 2002 IEEE International Conference on Fuzzy Systems. FUZZ-IEEE 2002. Proceedings (Cat. No. 02CH37291), vol. 2, pp. 1162–1167 (2003). <https://doi.org/10.1109/fuzz.2002.1006668>
42. Haykin, S.: *Neural networks: a comprehensive foundation* (1999)
43. Wang, Y., Wang, Y., Witten, I.: Inducing model tree for continuous classes. In *Proceedings of Poster Papers, 9th European Conference on Machine Learning*, pp. 128–137 (1997)
44. Salzberg, S.L.: *C4.5: Programs for machine learning* by J. Ross Quinlan. *Mach. Learn.* **16**, 235–240. Morgan Kaufmann Publishers, Inc., 1993 (1994). <https://doi.org/10.1007/BF00993309>
45. Idri, A., El Ouassif, B., Hosnia, M., Abran, A.: Classification techniques in breast cancer diagnosis: a systematic literature review. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* (2020)
46. Schapire, E., R.: *A brief introduction to boosting* (1999)
47. Sergios, T., Konstantinos, K.: *Pattern Recognition, Third Edition*.
48. Polikar, R.: *Ensemble based systems in decision making* (2006). <https://doi.org/10.1109/MCAS.2006.1688199>
49. Ali, K., Michael J.P.: On the Link between Error Correlation and Error Reduction in Decision Tree Ensembles (1995)
50. Kuncheva, L.I., Whitaker, C.J.: Ten measures of diversity in classifier ensembles: limits for two classifiers. *IEE Colloq.* 73–82 (2001). <https://doi.org/10.1049/ic:20010105>
51. Udney Yule, G.: On the association of attributes in statistics: with illustrations from the material of the childhood society, & c on JSTOR. *Philos. Trans. R. Soc. London. A* **194**, 257–319 (63 pages) (1900)
52. Giacinto, G., Roli, F.: Design of effective neural network ensembles for image classification purposes. *Image Vis. Comput.* **19**, 699–707 (2001). [https://doi.org/10.1016/S0262-8856\(01\)00045-2](https://doi.org/10.1016/S0262-8856(01)00045-2)
53. Cunningham, P., Carney, J.: Diversity versus quality in classification ensembles based on feature selection. In: López de Mántaras, R., Plaza, E. (eds.) *ECML 2000. LNCS (LNAD)*, vol. 1810, pp. 109–116. Springer, Heidelberg (2000). [https://doi.org/10.1007/3-540-45164-1\\_12](https://doi.org/10.1007/3-540-45164-1_12)
54. Partridge, D., Krzanowski, W.: Software diversity: Practical statistics for its measurement and exploitation. *Inf. Softw. Technol.* **39**, 707–717 (1997). [https://doi.org/10.1016/s0950-5849\(97\)00023-2](https://doi.org/10.1016/s0950-5849(97)00023-2)
55. Banfield, R.E., Hall, L.O., Bowyer, K.W., Kegelmeyer, W.P.: A new ensemble diversity measure applied to thinning ensembles. In: Windeatt, T., Roli, F. (eds.) *MCS 2003. LNCS*, vol. 2709, pp. 306–316. Springer, Heidelberg (2003). [https://doi.org/10.1007/3-540-44938-8\\_31](https://doi.org/10.1007/3-540-44938-8_31)
56. Kadkhodaei, H., Moghadam, A.M.E.: An entropy based approach to find the best combination of the base classifiers in ensemble classifiers based on stack generalization. In: 2016 4th International Conference on Control, Instrumentation, and Automation, ICCIA 2016, pp. 425–429. Institute of Electrical and Electronics Engineers Inc. (2016). <https://doi.org/10.1109/ICCIAutom.2016.7483200>
57. Nascimento, D.S.C., Canuto, A.M.P., Silva, L.M.M., Coelho, A.L.V.: Combining different ways to generate diversity in bagging models: an evolutionary approach. In: *Proceedings of the International Joint Conference on Neural Networks*, pp. 2235–2242. IEEE (2011). <https://doi.org/10.1109/IJCNN.2011.6033507>
58. Lysiak, R., Kurzynski, M., Woloszynski, T.: Optimal selection of ensemble classifiers using measures of competence and diversity of base classifiers. *Neurocomputing* **126**, 29–35 (2014). <https://doi.org/10.1016/j.neucom.2013.01.052>

59. Lopes Bhering, L., Cruz, D., Soares De Vasconcelos, E., Ferreira, A., Fernando, M., De Resende, R.: Alternative methodology for Scott-Knott test. *Crop Breed. Appl. Biotechnol.* **8**, 9–16 (2008)
60. Cox, D.R., Spjøtvoll, E.: On partitioning means into groups source. Wiley behalf Board Found. *Scand. J. St.* **9**, 147–152 (1982)
61. Calinski, T., Corsten, L.C.A.: Clustering means in ANOVA by simultaneous testing. *Biometrics* **41**, 39 (1985). <https://doi.org/10.2307/2530641>
62. Sharma, A., Kulshrestha, S., Daniel, S.: Machine learning approaches for breast cancer diagnosis and prognosis. In: 2017 International Conference on Soft Computing and its Engineering Applications: Harnessing Soft Computing Techniques for Smart and Better World, *icSoft-Comp 2017*, pp. 1–5. Changa, India (2018). <https://doi.org/10.1109/ICSOFTCOMP.2017.8280082>
63. Bony, S., Pichon, N., Ravel, C., Durixl, A., Balfourier, F.: The relationship between mycotoxin synthesis and isolatemorphology in fungal endophytes of *Lolium perenne*. *New Phytol.* **152**, 125–137 (2001)
64. Tsoumakas, G., Angelis, L., Vlahavas, I.: Selective Fusion of Heterogeneous Classifiers. *Intell. Data Anal.* **9**, 511–525 (2005). <https://doi.org/10.3233/ida-2005-9602>
65. Borges, L., Ferreira, D.: Power and type I errors rate of Scott-Knott, Tukey and Newman-Keuls tests under normal and no-normal distributions of the residues. *Rev. Matemática e Estatística.* **21**, 67–83 (2003)
66. Rowley, C.K.: Borda, Jean-Charles de (1733–1799). In: Durlauf, S.N., Blume, L.E. (eds.) *The New Palgrave: Dictionary of Economics*, pp. 527–529. Palgrave Macmillan UK, London (2008). [https://doi.org/10.1007/978-1-349-58802-2\\_148](https://doi.org/10.1007/978-1-349-58802-2_148)
67. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique (2002)
68. Gu, S.: Generating diverse and accurate classifier ensembles using multi-objective optimization (2014)
69. WEKA-University of Waikato: WEKA. <https://ai.waikato.ac.nz/weka/>
70. Smith, B.L., Scherer, W.T., Conklin, J.H.: Exploring imputation techniques for missing data in transportation management systems. *Transp. Res. Rec. J. Transp. Res. Board.* **1836**, 132–142 (2003). <https://doi.org/10.3141/1836-17>
71. Idri, A., Abnane, I., Abran, A.: Missing data techniques in analogy-based software development effort estimation. *J. Syst. Softw.* **117**, 595–611 (2016). <https://doi.org/10.1016/J.JSS.2016.04.058>
72. Oh, S.B.: On the relationship between majority vote accuracy and dependency in multiple classifier systems. *Pattern Recogn. Lett.* **24**, 359–363 (2003). [https://doi.org/10.1016/S0167-8655\(02\)00260-X](https://doi.org/10.1016/S0167-8655(02)00260-X)
73. Kuncheva, I.L.: *Combining Pattern Classifiers: Methods and Algorithms* (2014). <https://doi.org/10.1002/97811189145641>
74. Idri, A., Hosni, M., Abran, A.: Improved estimation of software development effort using classical and fuzzy analogy ensembles. *Appl. Soft Comput. J.* **49**, 990–1019 (2016). <https://doi.org/10.1016/j.asoc.2016.08.012>
75. Hansen, L.K., Salamon, P.: Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.* **12**, 993–1001 (1990). <https://doi.org/10.1109/34.58871>
76. Zhou, Z.H., Wu, J., Tang, W.: Ensembling neural networks: many could be better than all. *Artif. Intell.* **137**, 239–263 (2002). [https://doi.org/10.1016/S0004-3702\(02\)00190-X](https://doi.org/10.1016/S0004-3702(02)00190-X)