IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 July 2023, accepted 3 August 2023, date of publication 14 August 2023, date of current version 17 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3304912

RESEARCH ARTICLE

Dynamic Ensemble Algorithm Post-Selection Using Hardness-Aware Oracle

PAULO R. G. CORDEIRO[1,2], (Member, IEEE),
GEORGE D. C. CAVALCANTI[1], (Senior Member, IEEE),
AND RAFAEL M. O. CRUZ[3], (Member, IEEE)
[1]Centro de Informática, Federal University of Pernambuco, Recife 50740-560, Brazil
[2]Instituto Federal de Pernambuco, Campus Barreiros, Barreiros 55560-000, Brazil
[3]Software Engineering Department, École de Technologie Supérieure, Montreal, QC H3C 1K3, Canada

Corresponding author: Paulo R. G. Cordeiro (prgc@cin.ufpe.br)

ABSTRACT Dynamic Ensemble Selection (DES) algorithms have obtained better performance in many tasks compared to monolithic classifiers and static ensembles. However, it is reasonable to assume that no DES algorithm is the optimal solution in different scenarios since diversity plays an important role. Thus, this paper addresses this research gap by proposing a novel approach called Hardness-aware Oracle with Dynamic Ensemble Selection (HaO-DES) that operates as a post-selection strategy, evaluating and selecting the best DES techniques per instance. Each DES technique ensemble is evaluated using a new measure called Hardness-aware Oracle (HaO). HaO extends the traditional Oracle concept by assessing a DES technique based on how the classifiers in the selected ensemble work together, contrasting with the individual classifier evaluation in the traditional assessment. We performed experiments over 30 databases, using three base classifiers (Perceptron, Logistic Regression, and Naive Bayes) in homogeneous and heterogenous pools' configurations, to assess HaO-DES with four DES approaches (KNORA-U, KNOP, DES-P, and META-DES). We use three performance metrics to evaluate the experiments: accuracy, F-score, and Matthews Correlation Coefficient (MCC). The results show that our approach outperforms or obtains similar results against the four individual DES approaches, mainly when considering heterogeneous pool settings. We also demonstrated the HaO-DES efficiency in choosing suitable DES techniques in different situations.

INDEX TERMS Dynamic ensemble selection, multiple classifier systems, oracle, hardness-aware oracle.

I. INTRODUCTION

Multiple Classifier Systems (MCS) have been developed as a counterpoint to approaches that use individual classifiers, aiming to use multiple classifiers to improve the effectiveness of single classification systems. MCS is generally composed of three [1] phases: generation, selection, and combination.

The generation phase is responsible for training the base classifiers, thus creating a pool of classifiers. The classifiers in the pool should preferably be complementary or diverse, meaning they should disagree in their decisions. The approaches developed for generating a pool of classifiers use, for example, different distributions of the training datasets, such as Bagging [2], as well as different models and variations in the parameters of the learning algorithms. Pools formed using classifiers from the same learning algorithm are called homogeneous. Pools formed by different learning algorithms, such as Support Vector Machine (SVM), k-NN (k-Nearest Neighbors), and Decision Tree (DT), are called heterogeneous [3]. Heterogeneous ensembles tend to be more diverse than homogeneous ones because of their different mathematical formulations, which usually implies different classification results [4].

The selection phase aims to find, within a pool, a classifier or a subset of classifiers that are best suited for a given problem. This subset is called an ensemble. It can be either static or dynamic. Static is when the same classifiers are selected for the entire test set, and dynamic is when different classifiers are selected for different test samples [5]. There are two types of dynamic selection methods: Dynamic Classifier Selection

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

(DCS), when only a classifier is selected, or Dynamic Ensemble Selection (DES) when an ensemble is selected; the latter is the focus of this research. Dynamic selection techniques commonly use a criterion to guide the selection process, including [1]: meta-learning (e.g., META-DES), accuracy (e.g., Dynamic Ensemble Selection Performance (DES-P)), and Oracle (e.g., K-Nearest Oracles Eliminate (KNORA-E) and K-Nearest Oracles Union (KNORA-U)). The Oracle [6] is a theoretical model that chooses a classifier for each query sample if there is at least one classifier in the pool that can correctly predict the query's label. It means that this concept, when applied to the selection phase, does not evaluate situations where few competent classifiers (i.e., classifiers that correctly answer a sample query) are in the pool. Moreover, Oracle's evaluation is performed using global information, whereas dynamic selection techniques should focus on local information [7]. It is important to note that the selection phase is not mandatory. For instance, Boosting and Random subspace [8] methods do not apply the selection phase. The last phase of an MCS is combination or integration. This phase aims to combine all classifiers in the ensemble if more than one classifier is chosen.

At present, the primary focus of research in DES is to develop novel approaches for enhancing its phases, including generation [9], selection [10], and combination [11]. Despite significant advancements, there is no guarantee that a single DES technique can be applied to resolve all problems. This statement aligns with the statistical reasoning behind MCS [12], which proposes that combining multiple classifiers enhances the likelihood of discovering the best solution for a given problem. Regarding the selection phase, research papers on Dynamic Ensemble Selection usually focus on finding ways to select base classifiers. However, the selection phase constructs an ensemble by individually selecting multiple classifiers without attempting to evaluate if the selected models perform well together or not. Furthermore, research papers are not concerned with assessing and selecting ensembles already generated by DES techniques.

Our research regarding papers that proposes combining multiple DES approaches found only one work by Elmi and Eftekhari [10], where they presented the Multi-Layer Selector (MLS). MLS applies a Bootstrap Sampling method in the generation phase, as in Bagging. In the selection phase, MLS works on multiple layers that employ DES methods to select suitable classifiers and transmit them to the subsequent layer. The DES methods are integrated within the MLS framework, allowing for collaborative efforts and interaction to eliminate unsuitable classifiers at each layer by combining the selection phase criteria. In combination, the majority vote is applied to integrate the selected classifiers. However, MLS does not allow the evaluation of all ensembles generated by DES methods since the selection evaluates one classifier at a time. As per the author's knowledge, the DES field still lacks methodologies to evaluate the ensembles selected by a set of DES techniques and investigate the benefits of pre-selected ensembles in achieving improved performance.

Therefore, the present work addresses this research gap in DES by proposing: "How can we evaluate and choose ensembles selected from a set of different DES techniques to improve the performance of these techniques used individually?" To investigate this question, we propose a novel approach called Hardness-aware Oracle with Dynamic Ensemble Selection (HaO-DES). HaO-DES is based on the assumption that different selection criteria can result in diverse selected ensembles, and the optimal criteria for selecting an ensemble may vary depending on the instance. Our proposal operates as a post-selection strategy that evaluates and selects the technique from a set of DES defined by the user, to generate more reliable predictions.

Additionally, our proposal helps by evaluating the whole ensemble instead of each model separately, as DES techniques usually work. Given the absence of a quantifiable means to assess diverse ensembles chosen through a collection of DES techniques, we present the concept of a Hardness-aware Oracle (*HaO*). *HaO* represents a novel evaluation metric inspired by the Oracle. It enhances the original oracle's concept by analyzing not only the existence of a single competent classifier but also the capacity of a DES technique to find a significant number of such classifiers.

In summary, this paper presents two contributions to the area of Dynamic Ensemble Selection, which are the following:

1) A new method that works as a post-selection scheme for dynamically selecting one ensemble from several ensembles given a set of DES techniques using Hardness-aware Oracle.
2) A novel metric to evaluate distinct ensembles selected from a set of DES techniques called the Hardness-aware Oracle (*HaO*), derived from the Oracle concept. The *HaO*, presented in section **Hardness-aware Oracle**, evaluates how good a DES technique is regarding the number of competent classifiers selected.

The paper is organized as follows: Section **Literature Review** discusses concepts related to MCS, especially DES, and presents research using heterogeneous and homogeneous ensembles, section **Hardness-aware Oracle** presents Hardness-aware Oracle, Section **Hardness-aware Oracle with Dynamic Ensemble Selection** details the proposed approach of this paper (HaO-DES), Section **Experimental Methodology** presents the experiments' methodology and Section **Results and Discussion** shows results. Finally, Section **Conclusion** presents the final discussions of the research, as well as possible directions for future works.

II. LITERATURE REVIEW

Multiple Classifier Systems (MCS) is a very active research field that aims to improve classification performance based on the assumption that using several classifiers may achieve better results when compared to situations where only one classifier is used [13]. An MCS is usually developed in three phases: (i) generation, where a pool of classifiers is generated; (ii) selection, where a subset of the classifiers (ensemble) created in the generation phase will be chosen; and (iii) integration, which aims to combine ensembles to classify a given query sample.

In the generation phase, a pool of classifiers $P = \{C_1, C_2, \ldots, C_m\}$ with m classifiers C is created. Those classifiers should be both diverse and accurate. Diversity means that classifiers should not make the same prediction mistakes. Such properties are important in order to cover the feature space properly. Even been a key point in MCS research, there is no formal definition to calculate diversity between two classifiers. However, some measures have been proposed to calculate diversity in a pool, such as [14]: Q-statistic, disagreement measure, and double-fault measure. Several approaches can be implemented to generate a pool, such as [1]: different distributions of the training set, namely, Bagging [2]; different parameters for the same base classifier, e.g., variations in the number of neighbors in the k-Nearest Neighbors (k-NN) or different base classifiers, called heterogeneous ensembles [3]. Heterogeneous ensembles tend to be more diverse than homogeneous ones because of their different mathematical formulations, which usually implies different classification results [4]. We survey some research that uses homogeneous and heterogeneous ensembles in its generation phase to investigate which classifiers are used and how the selection phase works. Table 1 presents the acronyms of the classifiers used in research papers that were not yet cited in this text. Information about works that use homogeneous pools is presented in Table 2, and the works that use heterogeneous *pools* are listed in Table 3.

In the second phase, selection, the objective is to a subset of classifiers ($P' \subseteq P$), also called an ensemble of classifiers. There are two approaches: static and dynamic [1]. In the static approach, a subset of the classifiers is chosen for all test samples. In contrast, in the dynamic approach, called Dynamic Ensemble Selection (DES), a subset of the pool is chosen for each query sample. In dynamic selection, the classifier selection is performed using some criteria, given the pool created in the previous phase. Among the criteria found in the literature, one can mention ranking (*e.g.* DSC-Rank); Oracle, as is the case of KNORA-E, KNORA-U [5], and K-Nearest Output Profile (KNOP); accuracy, such as DES Performance (DES-P) [15]; and meta-learning [16], as is the case of META-DES. The criteria presented are commonly computed from the Region of Competence (RoC) denoted as $\theta_{\mathbf{x}_q}$, a fundamental concept for dynamic selection approaches. RoC is a local region calculated for a query sample, such that $\theta_{\mathbf{x}_q} = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$, where k is the size of the ROC. It is usually obtained by applying k-NN or clustering methods on a validation set (DSEL) or the training set itself. The last phase in an MCS is integration, also called aggregation or combination, whose goal is to combine the classifiers selected in the selection phase when more than one classifier is selected. Examples of techniques from this phase are majority vote, product rule, and sum rule [13].

TABLE 1. List of classifiers and their respective acronyms used in the heterogeneous ensemble research.

Classifier	Acronyms
Adacost	ADAC
Bagging	BAG
Cost-Sensitive CART	C-CART
Cost-Sensitive Random Forest	C-RF
Decision Table	DET
Decision Stump	DS
Deep Neural Network	DNN
Fisher Classifier	FC
Gaussian Bayes	GB
Gaussian Process Classifier	GPC
KStar	k*
Linear Discriminant Analysis	LDA
Linear Discriminant Classifier	LDC
Logistic Classifier	LC
Logistic Regression	LR
Multilayer perceptron	MLP
Nearest Mean Classifier	NCM
Hoeffding Tree	HT
Naive Bayes	NB
OneR	1R
PART Rule Learning Algorithm	PART
Quadritic Discriminant Analysis	QDA
Radial Basis Function Network	RBF
Random Forests	RF
Random Subspace Method	RSM
Reptree	RT
Rotation Forest	RTF
Stochastic Gradient Boosting	SGB
RIPPER rule learning	JRIP
Sequential Minimal Optimization	SMO
Support Vector Machine with Quadratic Kernel	SVMQ
Trimmed Bagging	TBAG

TABLE 2. Summary of information (year, base classifiers, and selection methods) related to research papers on Multiple Classifier Systems using homogeneous ensembles.

Research	Year	Classifier	Selection
WOLOSZYNSKI et al., [17]	2011	DT	DES
WOLOSZYNSKI et al., [15]	2012	DT	DES
CRUZ et al., [16]	2015	Perceptron	DES
GUO et al., [18]	2018	DT	Static
NGUYEN et al., [19]	2019	DT	Static
WANG et al., [20]	2021	DT	Static
ELMI et al., [10]	2021	Perceptron	DES
MOHAMMED et al., [21]	2022	DT	Static

Analyzing Tables 2 and 3, one can notice several types of base classifiers used by the researchers. For the homogeneous approaches, unstable classifiers are more common (e.g., DT and Perceptron). In contrast, for the approaches using heterogeneous pools, we have linear (e.g., LR) and non-linear (e.g., k-NN and SVM) classifiers. However, according to the authors' research, no study has examined heterogeneous pools using exclusively simple algorithms with the following characteristics: diversity and low computational cost, such as NB, LR, and Perceptron.

TABLE 3. Summary of information (year, base classifiers, and selection methods) related to research papers on Multiple Classifier Systems using heterogeneous ensembles.

Research	Year	Classifiers	Selection
TSOUMAKAS et al., [3]	2005	DET, Jrip, PART, DT, k-NN, K*, NB, SMO, RBF, MLP	Static
WOLOSZYNSKI et al., [17]	2011	LC, NMC, k-NN, Parzen, DT, SVM	DES
WOLOSZYNSKI et al., [15]	2012	LDC, NMC, k-NN, Parzen, DT, MLP	DES
NANI et al., [22]	2015	SVM, GPC, AdaBoost	No
HAQUE et al., [23]	2016	NB, SVM, LC, k-NN, DET, DT, RF	No
LARGE et al., [24]	2017	LC, C4.5, SVM, k-NN, MLP, RF, RTF, SVMQ, DNN	No
COSTA et al., [11]	2018	k-NN, MLP, DT, NB, SVM	No
NGUYEN et al., [25]	2018	LDA, NB, k-NN, C.45, DS, FC, NCM, LC	No
FILHO et al., [26]	2018	MLP, NB, DT, k-NN, SVM	DES
BOCK et al., [27]	2020	BAG, TBAG, SGB, RTF, RSM, RF, CART, C4.5, C4.4, LR, LDA, QDA, MLP, SVM, k-NN, ADAC, C4.5 + MC, C-RF, C-CART	Static
KADKHODAEI et al., [28]	2020	NB, BN, K*, k-NN, DT, RT, DS, 1R, DET, SVM	Static
OSTVAR; MOGHADAM, [29]	2020	NB, BN, K*, k-NN, DT, RT, DS, 1R, DET, SVM	DES
ZYBLEWSKI et al., [30]	2021	GB, HT, k-NN, SVM	DES
WANG et al., [4]	2021	GB, LR, QDA, k-NN, DT, RF, XGBoost	DES

A. THE ORACLE

Kuncheva introduced the Oracle [6] as a conceptual model that chooses a classifier capable of providing a correct response to a query sample, given the presence of at least one classifier that correctly classifies it. Because it is the perfect selection model, the Oracle is regarded as a potential upper limit for DCS techniques.

Although abstract, the Oracle finds practical application in various phases of an MCS. For instance, the work [7] proposes to apply Oracle's concept in the generation phase by iteratively generating hyperplanes, using Perceptrons, to ensure that each instance in the training set is accurately classified by at least one of the base classifiers in the pool. The Oracle is also applied in the selection phase, exemplified by KNORA-U and KNORA-E [5]. To illustrate the application of the Oracle concept in the selection phase using KNORA-U: given a query sample $\mathbf{x}_q$, a pool P, and the RoC of $\mathbf{x}_q$ defined as $\theta_{\mathbf{x}_q}$, KNORA-U selects the classifiers that correctly at least one samples on $\theta_{\mathbf{x}_q}$.

III. HARDNESS-AWARE ORACLE

In Multiple Classifier Systems, a classifier selection technique can be considered optimal if it can find the competent classifiers if they exist (i.e., classifiers that correctly predict the instance presented to the system). As we discussed before, in the context of DES, researchers are concerned with proposing new methods to select ensembles by usually calculating the Region of Competence (RoC) for a query sample, then applying metrics such as accuracy or Oracle to guide the process to each model in the pool, thus generating an ensemble. Regarding the author's knowledge, there is no concern in evaluating the ensembles already generated by DES techniques to find the most suitable one for a given query sample.

Therefore, the Oracle [7] inspires us to propose a way to evaluate those ensembles, called Hardness-aware Oracle (*HaO*). However, Oracle is an abstract technique developed to theoretically assess whether at least one competent classifier is in the pool to classify an instance. Therefore, when the Oracle is applied as a criterion, the evaluation becomes superficial since it is only focused on analyzing the presence of a single competent classifier. For a more realistic analysis, we propose a new look at the Oracle, which aims to assess ensembles generated by a set of selections technique by evaluating the effectiveness of the dynamic selection algorithm in selecting the highest number of competent classifiers with the number of competent classifiers in the pool.

Hardness-aware Oracle is the ratio between the number of competent classifiers selected by a DES technique and the total classifiers selected by the same technique for each query sample present in a data set. Given a pool (P), consisting of n classifiers C, such that $P = \{C_1, C_2, \ldots, C_n\}$, given that $P'_{\mathbf{x}}$ is the *ensemble* of classifiers selected by a dynamic selection technique for a query sample ($\mathbf{x}$), such that $P'_{\mathbf{x}} \subseteq P$; given that $O_{\mathbf{x}}$ is the set of competent classifiers related to the class of $\mathbf{x}$, predicted by DES techniques, present in *pool* P, such that $O_{\mathbf{x}} \subseteq P$, $HaO_{\mathbf{x}}$ is defined by the following equation:

$$HaO_{\mathbf{x}}(P'_{\mathbf{x}}, O_{\mathbf{x}}) = \frac{\left|P'_{\mathbf{x}} \cap O_{\mathbf{x}}\right|}{\left|P'_{\mathbf{x}}\right|} \tag{1}$$

Exemplifying: given a *pool* P with 5 classifiers, such that $P = \{C_1, C_2, \ldots, C_5\}$; given that for a query sample $\mathbf{x}_1$, the set of competent classifiers ($O_{\mathbf{x}_1}$) consists of 4 classifiers: $\left|O_{\mathbf{x}_1}\right| = 4$; given that a classifier selection technique returns a set $P'_{\mathbf{x}_1}$ with 4 classifiers, $\left|P'_{\mathbf{x}_1}\right| = 4$, with 3 of them being competent, $\left|P'_{\mathbf{x}_1} \cap O_{\mathbf{x}_1}\right| = 3$, $HaO_{\mathbf{x}_1}$ is calculated as follows:

$$HaO_{\mathbf{x}_1}(P'_{\mathbf{x}_1}, O_{\mathbf{x}_1}) = \frac{\left|P'_{\mathbf{x}_1} \cap O_{\mathbf{x}_1}\right|}{\left|P'_{\mathbf{x}_1}\right|} = \frac{3}{4} = 0.75 \tag{2}$$

However, if for the same query sample $\mathbf{x}_1$ and for the same pool P, another classifier selection technique would result in a set of classifiers $P''_{\mathbf{x}1}$, with 4 competent classifiers, $\left|P''_{\mathbf{x}1} \cap O_{\mathbf{x}_1}\right| = 4$, calculate $HaO'_{\mathbf{x}1}$ as:

$$HaO'_{\mathbf{x}_1}(P''_{\mathbf{x}_1}, O_{\mathbf{x}_1}) = \frac{\left|P''_{\mathbf{x}_1} \cap O_{\mathbf{x}_1}\right|}{\left|P''_{\mathbf{x}_1}\right|} = \frac{4}{4} = 1 \tag{3}$$

Even if both techniques select some competent classifiers for the presented query sample $\mathbf{x}_1$, as $HaO'_{\mathbf{x}_1}(P'_{\mathbf{x}_1}, O_{\mathbf{x}_1}) < HaO'_{\mathbf{x}_1}(P''_{\mathbf{x}_1}, O_{\mathbf{x}_1})$, the second selection technique, which selects $P''_{\mathbf{x}_1}$, is more effective at finding every one of them.

IV. HARDNESS-AWARE ORACLE WITH DYNAMIC ENSEMBLE SELECTION

This section describes the new DES approach called Hardness-aware Oracle with Dynamic Selection Ensemble (HaO-DES). This approach proposes two novelties: (i) the incorporation of several ensembles of classifiers, selected by different dynamic ensemble selection techniques, and (ii) the use of the Hardness-aware Oracle as a criterion to select the best ensemble, among those generated by DES techniques, aiming to find better results than using individual techniques. HaO-DES is composed of three phases: **(1)** pool generation and preparation of the DES techniques, **(2)** selection, and **(3)** combination.

A. HaO-DES - PHASE 1: POOL GENERATION AND DES' PREPARATION

This section describes the first phase of HaO-DES, which aims to generate a pool of classifiers and prepare the DES techniques. In order to create a more diverse pool, this phase applies a Bootstrap Sampling (Stage 1.1) and then trains the base classifiers (Stage 1.2). Following, all DES techniques are initialized (Stage 1.3). The three stages are detailed below:

Stage 1.1: Bootstrap sampling. The idea of applying Bootstrap Sampling, similar to the Bagging approach, is to generate diversity in the pools. For the dataset T, Bootstrap sampling is applied, with the number of bootstraps set as b, thus generating b new datasets. For example, applying Bootstrap Sampling with $b = 2$, for the dataset T, two training sets are generated, T^1 and T^2.

Stage 1.2: Pool generation. In this stage, each classifier C_i, such that $i = 1, 2, \ldots, m$, where m is the number of base classifiers, is trained using the training sets (bootstraps) generated in Stage 1.1. For the approach proposed in this paper, the classifiers should be heterogeneous. The suggestion of heterogeneous classifiers for this stage is based on their diversity. Various types of classifiers can be used in this stage, including Perceptron, LR, and NB. The output of this stage is the pool $P = \{C_1^1, C_1^2, \ldots, C_m^b\}$, generated from the training of each of the m classifiers using each bootstrap, totaling a pool with $b \times m$ classifiers. The next stages perform the dynamic ensemble selection and then the combination. Stages 1.1 and 1.2 are represented in Fig. 1.

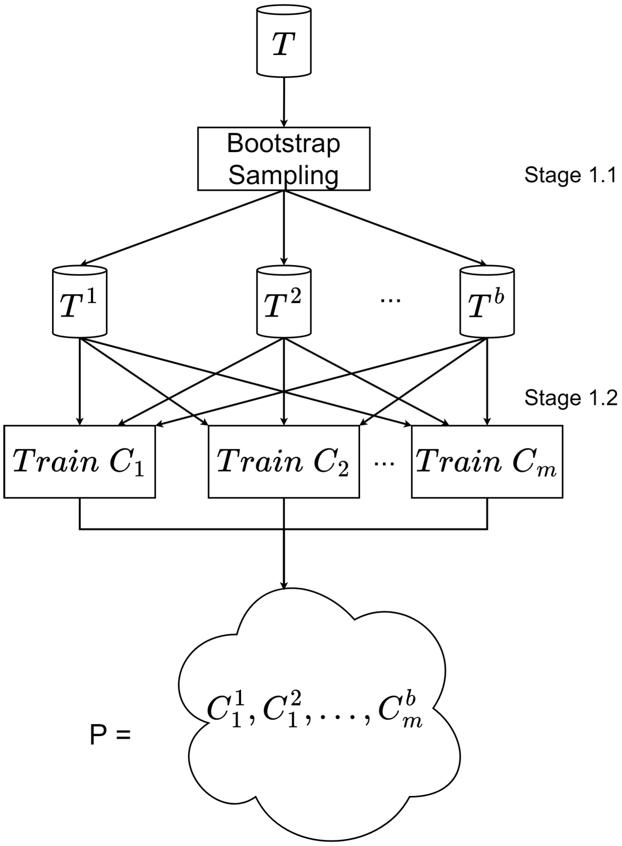


FIGURE 1. The initial two stages to the pool generation phase of HaO-DES. For the training set T, Bootstrap Sampling is applied, generating b new bootstraps. Each bootstrap is used to train one of the m base classifiers (C), thus forming the pool (P) with $b \times m$ classifiers.

Stage 1.3: Initialization and preparation of DES techniques. The goal of this stage is the initialization of the user-defined dynamic ensemble selection approaches (e.g., META-DES, KNORA-U, DES-P) in the DES_{set} set, containing n approaches, such that $DES_{set} = \{des_1, des_2, \ldots, des_n\}$, as well as their preparation. Each technique in the DES_{set} set is initialized from the pool, P, and subsequently prepared given the $DSEL$ validation data set. It is important to choose DES techniques with distinct selection criteria to increase the probability of selecting different ensembles.

B. HaO-DES - PHASES 2 E 3: DYNAMIC SELECTION AND COMBINATION

After the generation phase ends, Phase 2 begins, aiming to obtain an optimal ensemble dynamically given a set of DES techniques. Algorithm 1 summarizes how this phase works. For a query sample ($\mathbf{x_q}$), a validation set ($DSEL$), a pool

(P), and the set with Dynamic Ensemble Selection techniques (DES_{set}), the four stages in Phase 2 are specified as follows:

Stage 2.1: Definition of the Region of Competence ($\theta_{\mathbf{x}_q}$). Given the $\theta_{\mathbf{x}_q}$ and the validation set $DSEL$, the Region of Competence ($\theta_{\mathbf{x}_q}$) is defined using k-NN or clustering methods. The RoC is used for generating the ensembles selected by the DES techniques, and it is important to note that all DES techniques use the same RoC.

Stage 2.2: Ensemble generation by DES techniques. This stage's goal is to generate n ensemble (P^n). (P^n) is chosen from the dynamic selection phase of each DES technique from the DES_{set} for $\theta_{\mathbf{x}_q}$.

Stage 2.3: Hardness-aware Oracle Calculation. This stage performs the Hardness-aware Oracle (HaO) calculation for each of the P^n ensembles selected in Stage 2.2. Since we do not have the class label for $\mathbf{x}_q$, we assume that the output class for $\mathbf{x}_q$ is defined by the majority vote of the ensemble. For example, if we have a binary classification problem and an ensemble consisting of seven base classifiers denoted as $P' = \{C_1, \ldots, C_7\}$, selected using a specific DES technique, and let $y_{P'} = \{0, 1, 0, 0, 1, 1, 1\}$ represent the predictions of the base classifiers for $\mathbf{x}_q$, the majority vote of the ensemble would indicate class 1 as the answer. As stated in Stage 1.3, choosing DES techniques with different selection criteria increases the probability of selecting distinct ensembles, and hence, even with the same assumed class defined by the majority vote, they can have different values of HaO. For the query sample $\mathbf{x_q}$, pool P, and the ensemble P^n, the HaO is calculated according to Eq. 1.

Stage 2.4: Ensemble Selection. This stage is responsible for selecting the ensemble (P^{sel}) that obtained the highest HaO calculated in Stage 2.3.

After selecting the P^{sel}, Phase 3 begins, responsible for combining the classifiers into P^{sel}. Phase 3 consists of a single stage, described below:

Stage 3.1: Combination. This stage aims to combine the classifiers, if more than one has been selected, from the ensemble P^{sel} selected in Phase 2 to perform the classification. Techniques such as majority vote or sum rule can be proposed for this purpose.

V. EXPERIMENTAL METHODOLOGY

This section describes the experimental methodology used to compare the HaO-DES performance against individual DES approaches available in the literature. The analysis is conducted based on the Accuracy, F-score, and Matthews Correlation Coefficient (MCC) performance metrics. The experiments were implemented using Python 3.8.5, and combined with libraries Scikit-learn 1.0.1 [31] machine learning library and the DESlib 0.3 [32].

A. DATASET'S SETTINGS

Initially, the dataset T is split 50 % for training, 25 % for $DSEL$, and 25 % for test. The split is stratified, meaning that the proportions of the classes between training and testing are maintained. Then, we scaled the data using the Standard Scaler (i.e., Z-score normalization). The experiments were performed using 30 datasets from the UCI Machine Learning Repository [33], with different characteristics regarding the numbers of samples, dimensions, classes, and Imbalance Ratio (IR). The main characteristic of each dataset is presented in Table 4. The IR is calculated by the ratio between the cardinalities of the more numerous class by the less numerous one. We ran 30 experiment replications for each dataset, changing the distribution of the sets (holdout) to obtain the average values for the evaluated metrics.

TABLE 4. Datasets' description. The number of samples, dimensions (Dim), classes, and Imbalance Ratio (IR).

Datasets	Examples	Dim	Classes	IR
appendicitis	106	7	2	2.52
australian	690	14	2	1.12
balance	625	4	3	2.63
banana	5300	2	2	1.00
blood	748	4	2	3.20
bupa	345	6	2	1.38
cleveland	297	13	5	12.31
cmc	1473	9	3	1.30
column_3C	310	6	3	2.50
contraceptive	1473	9	3	1.88
dermatology	358	34	6	5.55
diabetes	768	8	2	1.86
glass1	214	9	2	1.82
glass6	214	9	2	6.38
german	1000	20	2	3.33
haberman	306	3	2	2.78
hayes-Roth	160	4	3	3.40
heart	270	13	2	1.25
ilpd	583	10	2	2.50
ionosphere	351	33	2	1.39
led7digit	500	7	10	1.54
mammographic	830	5	2	1.15
musk	476	166	2	1.29
pima	768	8	2	1.90
sonar	208	60	2	1.14
vehicle	846	18	4	1.10
vehicle2	846	18	2	2.88
vowel	990	13	11	1.00
wdbc	683	9	2	1.85
wine	178	13	3	1.00

B. HaO-DES' SETTINGS

The following paragraphs present the settings we used for Phases 1 and 3 of HaO-DES. Phase 2 does not require any setting configuration.

- HaO-DES - Phase 1: in Stage 1.1, Bootstrap Sampling, all the experiments used 100 bootstraps, similar to previous works in the literature [1], [5], [16]. For Stage 1.2, Pool Generation, three base classifiers were chosen to perform the experiments: Perceptron, Logistic Regression, and Naive Bayes. We choose those classifiers because they have distinct mathematical foundations and have a low computational cost [1], [11], [34]. Hence, being an appropriate set of learning algorithms for achieving a lightweight and diverse pool of classifiers.

Algorithm 1 HaO-DES Selection

```
procedure HaO_DES_Selection($\mathbf{x_q}$, $DSEL$, $P$, $DES_{set}$)
    $HaO_{max} \leftarrow 0$
    $P^{sel} \leftarrow \emptyset$
    $\theta_{\mathbf{x}_q} \leftarrow calculate\_ROC(\mathbf{x}_q, DSEL)$                ▷ (Stage 2.1)
    for $des$ in $DES_{set}$ do
        $P' \leftarrow get\_ensemble(des, \theta_{\mathbf{x}_q})$                  ▷ (Stage 2.2)
        $HaO_{des} \leftarrow calculate\_HaO(\mathbf{x}_q, P, P')$              ▷ (Stage 2.3)
        if $HaO_{des} \geq HaO_{max}$ then                                   ▷ (Stage 2.4)
            $HaO_{max} \leftarrow HaO_{des}$
            $P^{sel} \leftarrow P'$
        end if
    end for
    return $P^{sel}$
end procedure
```

As the research focuses not on optimizing each base model's hyperparameters, the default hyperparameters values from scikit-learn were used. Thus, the classifier pool (P) has 300 classifiers (3 baseline classifiers × 100 bootstraps). In Stage 1.3, four DES approaches were used: KNORA-U, KNOP, DES-P, and META-DES. These approaches were chosen because they feature distinct selection concepts (Oracle, accuracy, meta-learning) and are among the best-performing techniques according to a recent empirical study [1]. They were set using the default hyperparameter configurations of the DESlib 0.3 library. In order to make a fair comparison, the DES techniques use the same pool of classifiers and the same parameters for the base classifiers.

- HaO-DES - Phase 3: in Stage 3.1 (combination), majority voting is used, as it is a classifier combination approach widely used in the field.

C. PERFORMANCE EVALUATION METRICS

Three metrics are used to evaluate the techniques: accuracy (Eq. 4), which, although it is not a metric capable of effectively evaluating datasets that are imbalanced (i.e., high IR), is commonly used in the evaluation of DES techniques; F-score (Eq. 7) and Matthews Correlation Coefficient (MCC) (Eq. 8), which are interesting evaluative tools for datasets that have IR > 1, thus allowing a more precise evaluation of the experiment. For binary datasets, the metrics are defined according to the following equations:

$$\text{Accuracy} = \frac{TP + TF}{TP + TF + FP + FN} \tag{4}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{FN + TP} \tag{6}$$

TABLE 5. Relationship between a pair of classifiers.

	C_k correct (1)	C_k wrong (0)
C_i correct (1)	N^{11}	N^{10}
C_i wrong (0)	N^{01}	N^{00}

TABLE 6. Evaluation of *p-value* result of applying the Friedman's test for the comparison of performance HaO-DES and four DES approaches (KNOP, META-DES, DES-P, and KNORA-U), using three different base classifiers (Logistic Regression, Naive Bayes, and Perceptron) and three performance metrics (Accuracy (acc), F-score and Matthews Correlation Coefficient).

Base classifier	Metric	p-value
Logistic regression	acc	0.001
	F-score	0.000
	MCC	0.000
Naive Bayes	acc	0.001
	F-score	0.000
	MCC	0.000
Perceptron	acc	0.000
	F-score	0.000
	MCC	0.023

$$\text{F-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP) + (TN + FN)}} \tag{8}$$

where TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative. In cases of datasets with multiple classes, the $F - score_{micro}$ is calculated (Eq. 11). This requires $Precision_{micro}$ (Eq. 9) and $Recall_{micro}$ (Eq. 10), which are calculated individually for each class G, as follows:

$$Precision_{micro} = \frac{\sum_{i=1}^{G} TP_i}{\sum_{i=1}^{G} (TP_i + FP_i)} \tag{9}$$

$$Recall_{micro} = \frac{\sum_{i=1}^{G} TP_i}{\sum_{i=1}^{G} (FN_i + TP_i)} \tag{10}$$

TABLE 7. Evaluation of the average ranking, and *p-value* result of applying the paired Wilcoxon statistical test for the comparison of HaO-DES with four DES techniques: KNOP, META-DES, DES-P, and KNORA-U (KNRU). The pool is composed of 300 LR as the base classifiers. Accuracy, F-score, and MCC are evaluated.

Metric	Criterion	KNOP	META-DES	DES-P	KNRU	HaO-DES
ACC	ranking	3.48	2.62	3.14	3.55	**2.21**
	p-value	0.006	0.160	0.011	0.008	-
F-score	ranking	3.62	2.52	2.81	4.00	**2.05**
	p-value	0.001	0.069	0.002	0.001	-
MCC	ranking	3.57	2.57	3.00	3.76	**2.10**
	p-value	0.002	0.052	0.004	0.001	-

TABLE 8. Evaluation of the average ranking, and *p-value* result of applying the paired Wilcoxon statistical test for the comparison of HaO-DES with four DES techniques: KNOP, META-DES, DES-P, and KNORA-U (KNRU). The pool is composed of 300 NB as the base classifiers. Accuracy, F-score, and MCC are evaluated.

Metric	Criterion	KNOP	META-DES	DES-P	KNRU	HaO-DES
ACC	ranking	3.36	2.64	3.07	3.95	**1.98**
	p-value	0.004	0.048	0.028	0.001	-
F-score	ranking	3.52	2.81	2.76	3.76	**2.14**
	p-value	0.007	0.027	0.030	0.000	-
MCC	ranking	3.14	3.00	3.05	3.33	**2.48**
	p-value	0.045	0.025	0.108	0.032	-

TABLE 9. Evaluation of the average ranking, and *p-value* result of applying the paired Wilcoxon statistical test for the comparison of HaO-DES with four DES techniques: KNOP, META-DES, DES-P, and KNORA-U (KNRU). The pool is composed of 100 Perceptrons as the base classifiers. Accuracy, F-score, and MCC are evaluated.

Metric	Criterion	KNOP	META-DES	DES-P	KNRU	HaO-DES
ACC	ranking	3.41	2.86	2.62	3.81	**2.31**
	p-value	0.001	0.050	0.228	0.001	-
F-score	ranking	3.57	2.57	2.57	4.00	**2.29**
	p-value	0.000	0.040	0.038	0.000	-
MCC	ranking	3.62	2.81	2.29	4.05	**2.24**
	p-value	0.000	0.048	0.270	0.000	-

$$F - score_{micro} = 2 \times \frac{Precision_{micro} \times Recall_{micro}}{Precision_{micro} + Recall_{micro}} \quad (11)$$

While it is recognized that pairwise tests are generally preferred over multiple comparisons, as mentioned in [35], our objective is to present a range of viewpoints derived from the same experiments. Thus, we have employed two statistical analyses: the Wilcoxon signed ranks test for pairwise comparisons and the post-hoc Nemenyi's test for multiple comparisons, as suggested by [36].

D. DIVERSITY MEASURES

Additionally, we assess the diversity within the generated classifier pools using two measures [37]: disagreement (DIS) and double-fault (DF). These measures, applied to a pair of classifiers C_i and C_k based on the provided table of relationships (Table 5), are computed as follows:

$$DIS_{C_k,C_j} = \frac{N^{10} + N^{01}}{N^{11} + N^{10} + N^{01} + N^{00}} \quad (12)$$

$$DF_{C_k,C_j} = \frac{N^{00}}{N^{11} + N^{10} + N^{01} + N^{00}} \quad (13)$$

As those measures are pairwise diversity, we calculate an average measure for all pairs of classifiers in order to achieve a global metric using the following equation:

$$diversity_{average} = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{k=i+1}^{m} Q_{C_i,C_k} \quad (14)$$

where m is pool's size, and Q_{C_i,C_k} is a diversity pairwise measure calculated for classifiers C_i and C_k.

VI. RESULTS AND DISCUSSION

This section shows the results and discussions related to HaO-DES. We conducted experiments to evaluate four points. The first one is the effectiveness of Hardness-aware Oracle as a tool to select classifiers in homogeneous pools (**Section HaO-DES (homogeneous pools) analysis performance**) and heterogeneous pools (**Section HaO-DES (heterogeneous pools) analysis performance**). Both experiments compare our approach to four DES approaches: KNORA-U, KNOP, DES-P, and META-DES. The second point (**Section Heterogeneous vs. homogeneous pools: diversity**) aims to evaluate the diversity in homogeneous and heterogeneous pools by calculating two pairwise diversity measures (disagreement and double-fault) in order to validate if HaO-DES with homogeneous pools are more diverse than HaO-DES with heterogeneous. The third point (**Section HaO-DES: homogeneous vs. heterogeneous analysis performance**) compares the two pools' settings to analyze the

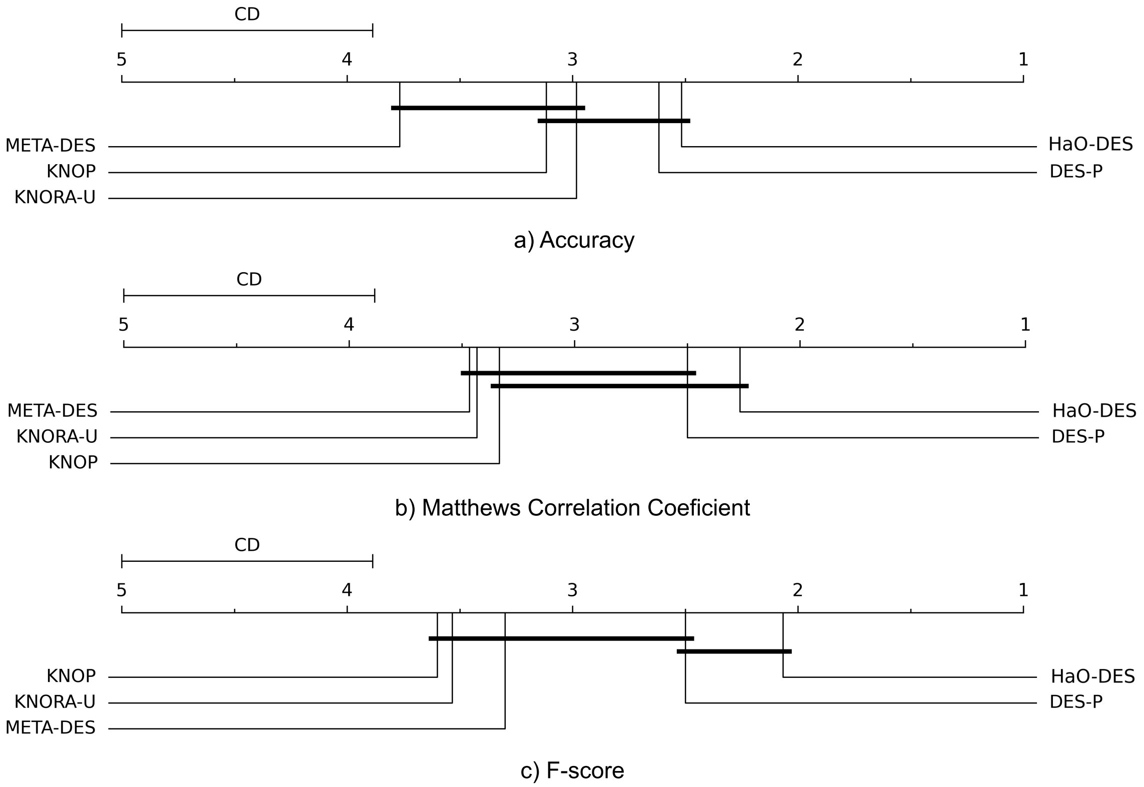


FIGURE 2. Critical Difference Diagram for Nemenyi's test for: a) accuracy, b) Matthews Correlation Coefficient, and c) F-score, evaluating HaO-DES against four DES techniques: DES-P, KNOP, KNORA-U, and META-DES. All approaches use a pool of three base classifiers (LR, NB, and Perceptron). The best approach is the one with the lowest rank.

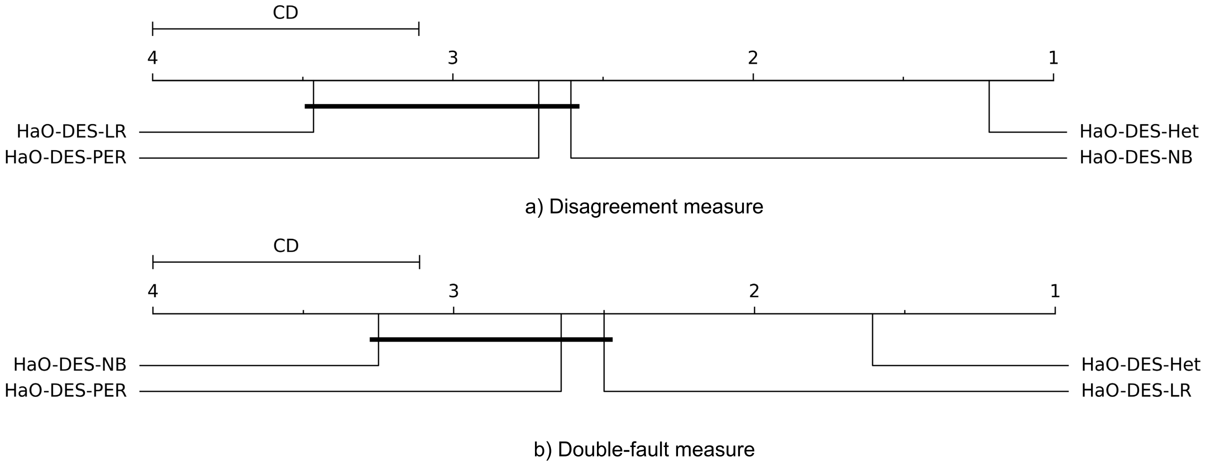


FIGURE 3. Critical Difference Diagram for Nemenyi's test for: a) Disagreement measure, and b) Double-fault measure, comparing HaO-DES with a heterogeneous pool (HaO-DES-Het) and three HaO-DES configurations with a homogeneous pool: LR (HaO-DES-LR), NB (HaO-DES-NB), and Perceptron (HaO-DES-PER).

impact a heterogeneous one has on the proposal. The fourth point (**Section Analyzing HaO-DES choices for DES techniques**) discusses the ability of HaO-DES to choose different DES for different situations. All statistical tests are performed with a confidence level of 95%. For multiple pairwise comparisons, we also performed the Bonferroni correction [35] for the p-values of the pairwise comparisons.

A. HaO-DES (HOMOGENEOUS POOLS) ANALYSIS PERFORMANCE

This section presents and discusses the performance of the HaO-DES setup with homogeneous pools using three different base classifiers: Logistic Regression, Naive Bayes, and Perceptron. HaO-DES was compared to four DES approaches (using the same pools): KNOP, META-DES, DES-P, and

TABLE 10. **Average accuracy evaluation comparing the HaO-DES approach to four DES techniques [16]: KNOP, META-DES (M-DES), DES-P, and KNORA-U (KNRU). All approaches use a heterogeneous pool of three base classifiers (LR, NB, and Perceptron). The best result from each dataset is presented in bold. The line (w/t/l) presents the number of wins, ties, and losses, and the ranking line presents the average ranking. The p-value(Wilcoxon) line presents the result of applying the paired Wilcoxon statistical test.**

Dataset	KNOP	M-DES	DES-P	KNRU	HaO-DES
appendicitis	0.860	0.859	**0.864**	**0.864**	0.863
australian	0.849	0.847	**0.850**	0.849	**0.850**
balance	0.887	**0.892**	0.883	0.881	0.891
banana	0.686	**0.743**	0.731	0.690	**0.743**
blood	0.778	0.776	0.784	0.779	**0.785**
bupa	**0.673**	0.643	0.665	0.667	0.668
cleveland	0.618	0.612	0.618	0.620	**0.621**
cmc	0.510	0.488	**0.513**	0.510	0.505
column_3C	**0.846**	0.840	0.845	0.845	0.838
contraceptive	**0.511**	0.487	0.510	0.507	0.506
dermatology	**0.972**	0.970	**0.972**	**0.972**	0.971
diabetes	**0.769**	0.763	0.764	**0.769**	0.766
glass1	0.645	0.667	0.677	0.643	**0.681**
glass6	0.938	**0.950**	0.947	0.938	**0.950**
german	0.742	0.739	**0.744**	0.741	0.741
haberman	0.732	0.727	**0.736**	0.729	0.732
hayes	0.609	0.661	0.642	0.618	**0.685**
heart	0.837	0.837	0.840	0.839	**0.844**
ilpd	0.708	0.674	0.681	**0.709**	0.684
ionosphere	**0.884**	0.545	0.883	**0.884**	0.575
led7digit	0.723	0.712	0.722	**0.726**	0.723
mammographic	**0.830**	0.826	0.828	**0.830**	0.826
musk	0.780	0.790	0.792	0.780	**0.796**
pima	**0.769**	0.764	0.766	**0.769**	**0.769**
sonar	0.776	0.789	0.787	0.776	**0.800**
vehicle	0.752	**0.762**	0.749	0.748	**0.762**
vehicle2	0.948	**0.958**	0.951	0.949	0.956
vowel	0.964	**0.979**	0.966	0.965	**0.979**
wdbc	**0.970**	0.969	**0.970**	**0.970**	0.969
wine	0.973	0.973	**0.975**	**0.975**	0.974
w/t/l	17/2/11	24/3/3	17/0/13	16/1/13	-
ranking	3.12	3.77	2.62	2.98	**2.52**
p-value(Wilcoxon)	0.086	0.000	0.086	0.079	-

KNORA-U (KNRU). We have first applied Friedman's test to asses if all the approaches (the four DES approaches and HaO-DES) perform equally. Table 6 presents the results of Friedman's test.

Since all p-values results were less than 0.05, as shown in Table 6, we can conclude that the methods do not perform equally. Therefore, we have applied the Wilcoxon test to evaluate if there is a pairwise statistical difference between HaO-DES and the DES approaches. Experimental results for each base classifier are presented in Tables 7, 8, and 9. The tables show two results: the average ranking of the approaches and the *p-value* of the result of applying the paired Wilcoxon statistical test in comparison to the techniques with HaO-DES.

Analyzing Tables 7, 8, and 9 when a pool of homogeneous classifiers was used, HaO-DES obtained the lowest rank for all the evaluated metrics and the three different types of base classifiers compared to the other approaches. When assessing the result of the Wilcoxon test, HaO-DES outperforms KNOP and KNORA-U on all metrics when the pool is composed of LR or Perceptron and on accuracy and F-score when the pool is composed of NB. Our approach obtained similar results in the statistical evaluation regarding all configurations and metrics compared to META-DES. HaO-DES outperforms DES-P on all three metrics regarding the pool formed by LR and obtains similar statistical results on other setups. In brief, HaO-DES achieved better or equivalent results across all three metrics compared to the approaches evaluated in all homogeneous pool configurations.

B. HaO-DES (HETEROGENEOUS POOLS) ANALYSIS PERFORMANCE

This section presents and discusses the performance of HaO-DES using a heterogeneous pool of three base classifiers (LR, NB, and Perceptron) compared to DES approaches (KNOP, META-DES, DES-P, and KNORA-U) while using the same HaO-DES base classifier configuration.

We have first applied Friedman's test to asses if all the approaches (the four DES approaches and HaO-DES) perform equally. Since all p-values results were less than 0.05 for all three metrics (p-value = 0.008 for accuracy, p-value = 0.000 for F-score, and p-value = 0.007 for MCC), meaning that the methods do not perform equally, we applied the Wilcoxon test to evaluate if there is a pairwise statistical difference between HaO-DES and the state-of-the-art DES approaches. To perform the comparison, three metrics

TABLE 11. Average MCC evaluation comparing the HaO-DES approach to four DES techniques [16]: KNOP, META-DES (M-DES), DES-P, and KNORA-U (KNRU). All approaches use a heterogeneous pool of three base classifiers (LR, NB, and Perceptron). The best result from each dataset is presented in bold. The line (w/t/l) presents the number of wins, ties, and losses, and the ranking line presents the average ranking. The p-value(Wilcoxon) line presents the result of applying the paired Wilcoxon statistical test.

Datasets	KNOP	M-DES	DES-P	KNRU	HaO-DES
appendicitis	0.511	0.531	**0.562**	0.525	0.545
australian	0.697	0.692	**0.698**	0.696	**0.698**
balance	0.801	**0.810**	0.794	0.791	0.809
banana	0.422	**0.523**	0.497	0.430	0.517
blood	0.198	0.222	0.276	0.201	**0.279**
bupa	**0.318**	0.258	0.302	0.304	0.309
cleveland	0.351	0.345	0.352	0.353	**0.359**
cmc	0.230	0.196	**0.239**	0.229	0.225
column_3C	**0.756**	0.746	0.755	0.754	0.743
contraceptive	0.230	0.195	**0.234**	0.224	0.226
dermatology	**0.965**	0.963	**0.965**	**0.965**	0.964
diabetes	**0.468**	0.455	0.459	**0.468**	0.464
glass1	0.034	0.208	**0.233**	0.023	0.221
glass6	0.725	**0.783**	0.770	0.719	0.778
german	0.315	0.315	**0.327**	0.313	0.323
haberman	0.142	0.139	**0.188**	0.116	0.186
hayes	0.413	0.493	0.470	0.428	**0.527**
heart	0.674	0.673	0.681	0.678	**0.688**
ilpd	0.113	0.157	0.140	0.119	**0.166**
ionosphere	**0.750**	0.278	0.748	0.749	0.317
led7digit	0.694	0.683	0.692	**0.697**	0.694
mammographic	**0.660**	0.651	0.655	**0.660**	0.650
musk	0.550	0.571	0.575	0.549	**0.583**
pima	0.468	0.457	0.463	0.468	**0.470**
sonar	0.557	0.581	0.575	0.556	**0.603**
vehicle	0.674	**0.685**	0.670	0.669	**0.685**
vehicle2	0.864	**0.891**	0.873	0.866	0.886
vowel	0.773	**0.876**	0.791	0.783	0.872
wdbc	0.933	0.932	**0.934**	0.933	0.932
wine	0.961	0.961	**0.963**	**0.963**	0.962
w/t/l	21/1/8	24/0/6	20/0/10	25/0/5	-
ranking	3.33	3.47	2.50	3.43	**2.27**
p-value(Wilcoxon)	0.004	0.000	0.049	0.001	-

were used: (i) accuracy (Table 10), (ii) Matthews Correlation Coefficient (Table 11), and (iii) F-score (Table 12). The tables show the average value of the metrics related to each dataset (rows) and approach used (columns), amount of wins, ties, and losses (row w/t/l), average ranking, and p-value of the result of applying the paired Wilcoxon statistical test. In order to compare all approaches rankings across all databases, we calculate Nemenyi's test for paired samples [36], and present the Critical Difference Diagram (CD) considering accuracy (Fig. 2. a), MCC (Fig. 2.b), and F-score (Fig. 2.c). Regarding the CD diagram interpretation, the approaches that achieve lower ranks are considered more effective on the chosen metric. Furthermore, techniques connected by a black bar are deemed statistically equivalent.

Upon examining the accuracy results in Table 10, it can be observed that HaO-DES outperforms META-DES (p-value = 0.000) and obtains similar statistical results with the other approaches evaluating p-value from the Wilcoxon test. The CD diagram in Fig. 2.a) shows similar results.

Evaluating MCC, Table 11 shows that HaO-DES outperforms KNOP, META-DES, and KNORA-U approaches according to Wilcoxon's test and obtains similar statistical results with DES-P. In addition, the MCC's evaluation from the CD diagram (Fig. 2.b) demonstrates HaO-DES outperforms KNORA-U and META-DES.

Analyzing the F-score results presented in Table 12, it is evident that the HaO-DES approach demonstrates statistical superiority over META-DES, KNORA-U, and KNOP when considering the p-value for the pairwise comparisons. Furthermore, the results are the same in the CD analysis depicted in Fig. 2.c), HaO-DES outperforms META-DES, KNORA-U, and KNOP.

Therefore by evaluating the three proposed metrics, HaO-DES is effective compared to established techniques in the literature on datasets with different settings, presenting itself as a new alternative for Dynamic Ensemble Selection. It is important to remember that no hyperparameter tuning was performed on any algorithm. Had it been the case, the hyperparameters' variations could change the results. Regarding the proposal's computational cost: because all DES techniques share the same RoC definition, one of the bottlenecks in DES approaches, the computational cost increase should not be too significant.

C. HETEROGENEOUS VS. HOMOGENEOUS POOLS: DIVERSITY

This section compares pool diversity generated in Phase 1.2 for homogeneous and heterogeneous pool settings. Diversity was calculated using the disagreement and double-fault

TABLE 12. Average F-score evaluation comparing the HaO-DES approach to four DES techniques [16]: KNOP, META-DES (M-DES), DES-P, and KNORA-U (KNRU). All approaches use a heterogeneous pool of three base classifiers (LR, NB, and Perceptron). The best result from each dataset is presented in bold. The line (w/t/l) presents the number of wins, ties, and losses, and the ranking line presents the average ranking. The p-value(Wilcoxon) line presents the result of applying the paired Wilcoxon statistical test.

Datasets	KNOP	M-DES	DES-P	KNRU	HaO-DES
appendicitis	0.714	0.741	**0.755**	0.723	0.747
australian	0.846	0.844	0.847	0.846	**0.848**
balance	0.615	0.628	0.612	0.611	**0.632**
banana	0.628	0.711	0.698	0.634	**0.713**
blood	0.531	0.568	0.605	0.532	**0.608**
bupa	0.627	0.623	0.636	0.618	**0.642**
cleveland	0.338	0.345	0.348	0.350	**0.352**
cmc	0.475	0.454	**0.484**	0.475	0.474
column_3C	0.799	0.792	**0.800**	0.798	0.789
contraceptive	0.476	0.454	**0.481**	0.471	0.476
dermatology	**0.968**	0.966	0.967	**0.968**	0.967
diabetes	0.725	0.719	0.722	0.725	**0.726**
glass1	0.424	**0.576**	0.574	0.423	0.563
glass6	0.848	**0.884**	0.875	0.845	0.882
german	0.625	0.636	0.637	0.623	**0.641**
haberman	0.508	0.519	0.549	0.498	**0.553**
hayes	0.623	0.676	0.661	0.635	**0.702**
heart	0.833	0.833	0.836	0.835	**0.840**
ilpd	0.499	**0.569**	0.550	0.503	0.568
ionosphere	**0.871**	0.499	0.870	**0.871**	0.538
led7digit	0.715	0.706	0.715	0.718	**0.719**
mammographic	**0.829**	0.824	0.826	**0.829**	0.824
musk	0.771	0.783	0.784	0.771	**0.789**
pima	0.725	0.720	0.724	0.726	**0.728**
sonar	0.773	0.786	0.783	0.773	**0.796**
vehicle	0.745	**0.757**	0.741	0.740	**0.757**
vehicle2	0.931	**0.945**	0.936	0.932	0.942
vowel	0.881	**0.935**	0.890	0.886	0.933
wdbc	0.966	0.966	**0.967**	0.966	0.966
wine	0.974	0.974	**0.975**	**0.975**	0.974
w/t/l	27/0/3	25/1/6	25/0/5	26/0/4	-
ranking	3.60	3.30	2.50	3.53	2.07
p-value(Wilcoxon)	0.000	0.000	0.029	0.000	-

measures globally, as shown in Eq. 14. We have applied Friedman's test to asses if all the approaches (three HaO-DES configurations using a homogeneous pool and one with the heterogeneous settings) perform equally. Since all p-values results were less than 0.05 (p-value = 0.000 for disagreement and double-fault), meaning the methods do not perform equally, we applied the Wilcoxon test to evaluate if there is a significant statistical difference between them. The results are presented in Table 13.

Regarding the results presented in Table 13, the configuration with heterogeneous pools is more diverse than the approach with homogeneous pools in both metrics. Figures 3 a) and 3 b) show the CD diagram for the disagreement measure and the double-fault measure, respectively. The two CD diagrams also demonstrate that the heterogeneous approach is more diverse than the homogeneous one.

D. SECTION HaO-DES: HOMOGENEOUS VS. HETEROGENEOUS ANALYSIS PERFORMANCE

This section compares the effectiveness of HaO-DES using a heterogeneous pool (HaO-DES-Het) against three HaO-DES settings of homogeneous pools: LR (HaO-DES-LR), NB (HaO-DES-NB), and Perceptron (HaO-DES-Perceptron). Comparing homogeneous and heterogeneous pools involves evaluating the impact of the classifiers' diversity on the performance of HaO-DES. We have first applied Friedman's test to asses if all the approaches (the four HaO-DES configurations) perform equally. Since all p-values results were less than 0.05 for all three metrics (p-value = 0.000 for accuracy, p-value = 0.004 for F-score, and p-value = 0.006 for MCC), meaning the methods do not perform equally, we applied the Wilcoxon test.

Table 14 presents a pairwise analysis and evaluates two criteria: the average ranking of the approaches and the p-value of the result of applying the paired Wilcoxon statistical test for the evaluated approaches. The CD Diagram using the Nemenyi's test considering accuracy, MCC, and F-score are presented in Fig. 4 a), b), and c), respectively.

The pairwise statistical analysis (Table 14) indicates that most metrics and criteria evaluated show an advantage of the heterogeneous pool over approaches that use homogeneous pools. The exceptions are when comparing HaO-DES-Het with HaO-DES-LR based on accuracy and MCC and HaO-DES-Het with HaO-DES-NB based on F-score and MCC, where their performances were deemed statistically equivalent. Nevertheless, the analysis of the CD diagram

TABLE 13. Evaluation of the average ranking, and p-value of the result of applying the paired Wilcoxon statistical test for the comparison of HaO-DES with a heterogenous pool (HET) against three configurations of homogeneous pools, HaO-DES-LR (LR), HaO-DES-NB (NB), and HaO-DES-Perceptron (Perceptron). Disagreement measure and double-fault measure are evaluated.

Diversity measure	Criterion	LR	NB	Perceptron	HET
Disagreement	ranking	3.46	2.60	2.72	**1.22**
	p-value	0.000	0.000	0.000	-
Double-fault	ranking	2.50	3.25	2.65	**1.60**
	p-value	0.002	0.000	0.002	-

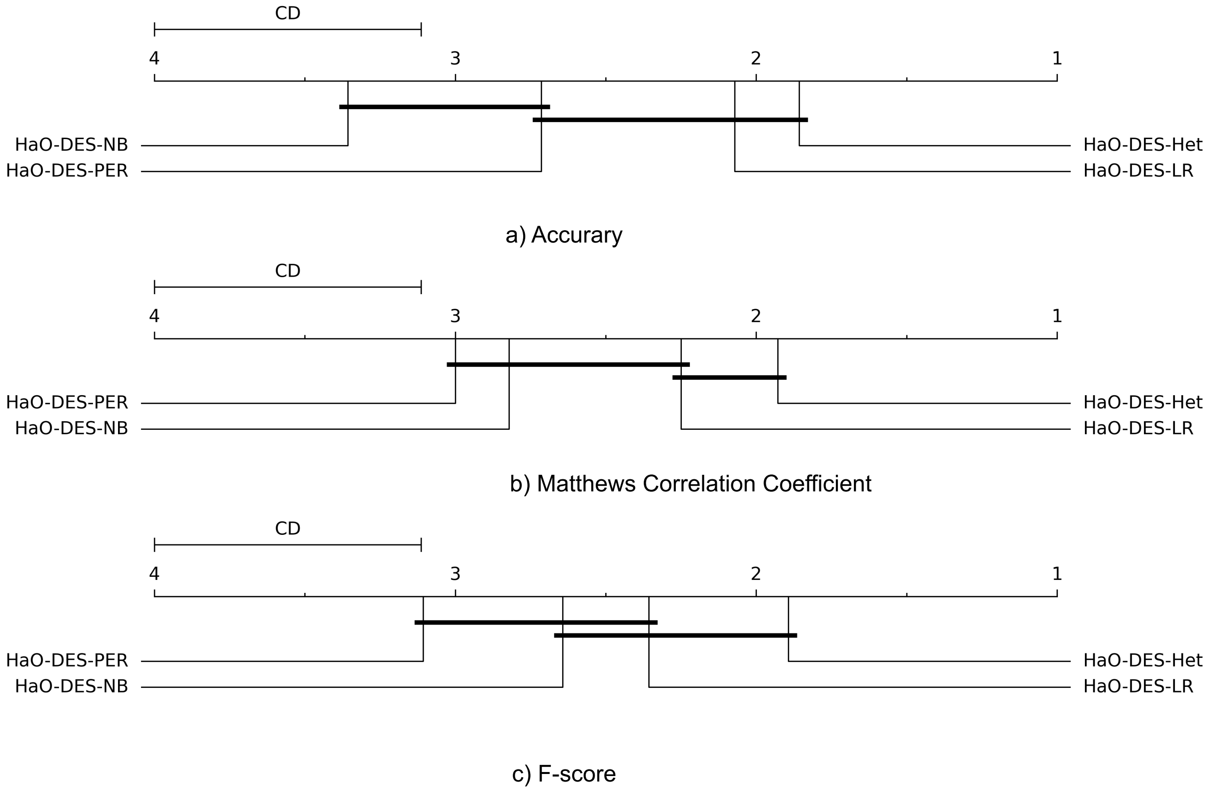


FIGURE 4. Critical Difference Diagram for Nemenyi's test for: a) accuracy, b) Matthews Correlation Coefficient, and c) F-score evaluating HaO-DES with a heterogeneous pool (HaO-DES-Het) and three HaO-DES configurations with a homogeneous pool: NB (HaO-DES-NB), LR (HaO-DES-LR), and Perceptron (HaO-DES-PER) and heterogeneous pools (HaO-DES-Het).

TABLE 14. Evaluation of the average ranking and p-value of the result of applying the paired Wilcoxon statistical test for the comparison of HaO-DES with a heterogenous pool (HET) against three configurations of homogeneous pools, HaO-DES-LR (LR), HaO-DES-NB (NB), and HaO-DES-Perceptron (Perceptron). Accuracy (ACC), F-score, and MCC are evaluated.

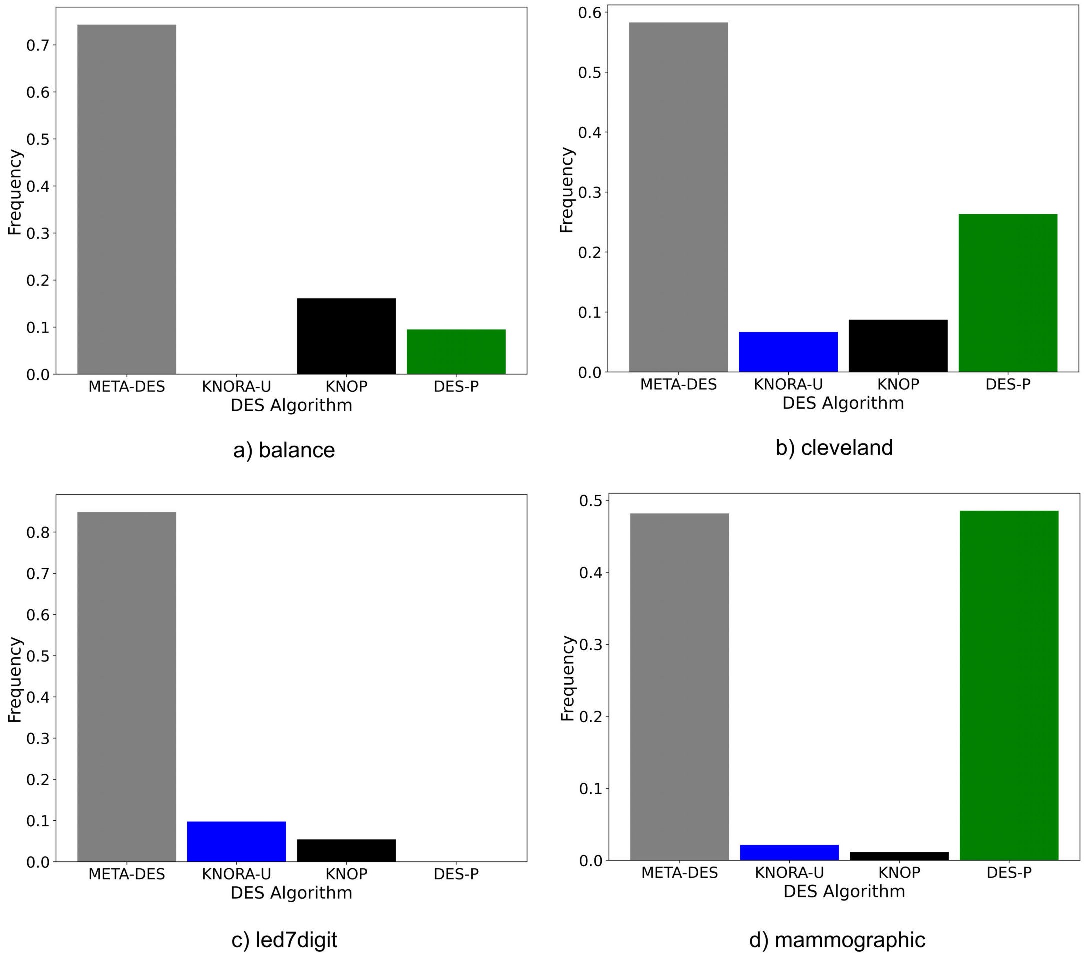
Metric	Criterion	LR	NB	Perceptron	HET
ACC	ranking	2.07	3.36	2.71	**1.86**
	p-value	0.132	0.000	0.001	-
F-score	ranking	2.36	2.64	3.11	**1.89**
	p-value	0.010	0.101	0.000	-
MCC	ranking	2.25	2.82	3.00	**1.93**
	p-value	0.048	0.020	0.000	-

(Fig. 4) indicates that HaO-DES-Het outperforms HaO-DES-NB in both accuracy and MCC evaluation and outperforms HaO-DES-PER on both MCC and F-score evaluation. In contrast, there was no statistical difference for the other settings of homogeneous pools. Such results suggest the ability of Hardness-aware Oracle to evaluate the best ensembles, especially when faced with the diversity generated by the heterogeneous classifiers, as shown in the previous section. The upcoming section discusses HaO-DES' behavior results in choosing diverse DES techniques.

E. ANALYZING HaO-DES CHOICES FOR DES TECHNIQUES

This last analysis aims to show the diversification of HaO-DES when selecting different ensembles (Stage 2.5) for different situations. We calculated how many times (%) HaO-DES selected the DES techniques for four datasets (Balance, Cleveland, Led7digit, and Mammographic), and the results are presented in Fig. 5.

Analyzing Fig. 5, we can see that HaO-DES selects different DES techniques for different situations, which can explain the better performance of HaO-DES compared to individual techniques in some datasets. META-DES is the most chosen technique at least 50% of the time, followed by DES-P. These results are unsurprising since they are the DES



a) balance

b) cleveland

c) led7digit

d) mammographic

FIGURE 5. Frequency of DES approach choices made by HaO-DES in Stage 2.5 for four databases (balance, cleveland, led7digit, and mammographic).

techniques with the best individual performance. However, depending on the problem, KNORA-U and KNOP are also chosen. This diversification of ensembles' choices makes HaO-DES effective on multiple datasets and improves performance compared to the best individual techniques.

VII. CONCLUSION

This work presented Hardness-aware Oracle with Dynamic Ensemble Selection (HaO-DES), a new Dynamic Ensemble Selection framework that uses several DES approaches to select different ensembles, and a new evaluation metric called Hardness-aware Oracle, used as a guide to distinguish which DES approaches selected the most appropriate ensemble. To analyze our proposal, we analyze two scenarios using a homogeneous and a heterogeneous pool. In both scenarios, HaO-DES achieves better or similar results when evaluating metrics such as accuracy, MCC, and F-score than the META-DES, KNOP, KNORA-U, and DES-P approaches. However, when using a more diverse pool, as in the case of heterogeneous scenarios, HaO-DES obtain better performance than homogeneous ones, suggesting its efficiency in selecting the best ensemble in this context. We also showed HaO-DES' capability to select appropriate DES techniques for different situations. Even when using the selection phase for all individual DES techniques, the computational cost is not a problem for HaO-DES because the Region of Competence, one of the bottlenecks in DES systems, is calculated only once for all query samples. Future works should focus on evaluating new classifiers and Dynamic Ensemble Selection techniques, analyzing the impact of the Multi-view Learning approach [38] on base classifier pool generation, and conducting a mathematical analysis of our proposed approach.

ACKNOWLEDGMENT

The authors would like to thank Instituto Federal de Pernambuco (IFPE), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo á Ciência e Tecnologia de Pernambuco (FACEPE), and the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

[1] R. M. O. Cruz, R. Sabourin, and G. D. C. Cavalcanti, "Dynamic classifier selection: Recent advances and perspectives," *Inf. Fusion*, vol. 41, pp. 195–216, May 2018.

[2] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[3] G. Tsoumakas, L. Angelis, and I. Vlahavas, "Selective fusion of heterogeneous classifiers," *Intell. Data Anal.*, vol. 9, no. 6, pp. 511–525, Dec. 2005.

[4] L. Wang, T. Mo, X. Wang, W. Chen, Q. He, X. Li, S. Zhang, R. Yang, J. Wu, X. Gu, J. Wei, P. Xie, L. Zhou, and X. Zhen, "A hierarchical fusion framework to integrate homogeneous and heterogeneous classifiers for medical decision-making," *Knowl.-Based Syst.*, vol. 212, Jan. 2021, Art. no. 106517.

[5] A. H. R. Ko, R. Sabourin, and A. S. Britto Jr., "From dynamic classifier selection to dynamic ensemble selection," *Pattern Recognit.*, vol. 41, no. 5, pp. 1718–1731, May 2008.

[6] L. I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 281–286, Feb. 2002.

[7] M. A. Souza, G. D. C. Cavalcanti, R. M. O. Cruz, and R. Sabourin, "On the characterization of the Oracle for dynamic classifier selection," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 332–339.

[8] M. Skurichina and R. P. W. Duin, "Bagging, boosting and the random subspace method for linear classifiers," *Pattern Anal. Appl.*, vol. 5, no. 2, pp. 121–135, Jun. 2002.

[9] M. Monteiro, A. S. Britto, J. P. Barddal, L. S. Oliveira, and R. Sabourin, "Exploring diversity in data complexity and classifier decision spaces for pool generation," *Inf. Fusion*, vol. 89, pp. 567–587, Jan. 2023.

[10] J. Elmi and M. Eftekhari, "Multi-layer Selector(MLS): Dynamic selection based on filtering some competence measures," *Appl. Soft Comput.*, vol. 104, Jun. 2021, Art. no. 107257.

[11] V. S. Costa, A. D. S. Farias, B. Bedregal, R. H. N. Santiago, and A. M. D. P. Canuto, "Combining multiple algorithms in classifier ensembles using generalized mixture functions," *Neurocomputing*, vol. 313, pp. 402–414, Nov. 2018.

[12] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, 2nd ed. Hoboken, NJ, USA: Wiley, 2014.

[13] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[14] C. A. Shipp and L. I. Kuncheva, "Relationships between combination methods and measures of diversity in combining classifiers," *Inf. Fusion*, vol. 3, no. 2, pp. 135–148, Jun. 2002.

[15] T. Woloszynski, M. Kurzynski, P. Podsiadlo, and G. W. Stachowiak, "A measure of competence based on random classification for dynamic ensemble selection," *Inf. Fusion*, vol. 13, no. 3, pp. 207–213, Jul. 2012.

[16] R. M. O. Cruz, R. Sabourin, G. D. C. Cavalcanti, and T. Ing Ren, "META-DES: A dynamic ensemble selection framework using meta-learning," *Pattern Recognit.*, vol. 48, no. 5, pp. 1925–1935, May 2015.

[17] T. Woloszynski and M. Kurzynski, "A probabilistic model of classifier competence for dynamic ensemble selection," *Pattern Recognit.*, vol. 44, nos. 10–11, pp. 2656–2668, Oct. 2011.

[18] H. Guo, H. Liu, R. Li, C. Wu, Y. Guo, and M. Xu, "Margin & diversity based ordering ensemble pruning," *Neurocomputing*, vol. 275, pp. 237–246, Jan. 2018.

[19] T. T. Nguyen, M. T. Dang, A. W. Liew, and J. C. Bezdek, "A weighted multiple classifier framework based on random projection," *Inf. Sci.*, vol. 490, pp. 36–58, Jul. 2019.

[20] Z. Wang, S. Zhao, Z. Li, H. Chen, C. Li, and Y. Shen, "Ensemble selection with joint spectral clustering and structural sparsity," *Pattern Recognit.*, vol. 119, Nov. 2021, Art. no. 108061.

[21] A. M. Mohammed, E. Onieva, M. Woźniak, and G. Martínez-Muñoz, "An analysis of heuristic metrics for classifier ensemble pruning based on ordered aggregation," *Pattern Recognit.*, vol. 124, Apr. 2022, Art. no. 108493.

[22] L. Nanni, S. Brahnam, S. Ghidoni, and A. Lumini, "Toward a general-purpose heterogeneous ensemble for pattern classification," *Comput. Intell. Neurosci.*, vol. 2015, pp. 1–10, Jan. 2015.

[23] M. N. Haque, N. Noman, R. Berretta, and P. Moscato, "Heterogeneous ensemble combination search using genetic algorithm for class imbalanced data classification," *PLoS ONE*, vol. 11, no. 1, Jan. 2016, Art. no. e0146116.

[24] J. Large, J. Lines, and A. Bagnall, "The heterogeneous ensembles of standard classification algorithms (HESCA): The whole is greater than the sum of its parts," 2017, *arXiv:1710.09220*.

[25] T. T. Nguyen, M. P. Nguyen, X. C. Pham, A. W.-C. Liew, and W. Pedrycz, "Combining heterogeneous classifiers via granular prototypes," *Appl. Soft Comput.*, vol. 73, pp. 795–815, Dec. 2018.

[26] J. A. S. L. Filho, A. M. P. Canuto, and R. H. N. Santiago, "Investigating the impact of selection criteria in dynamic ensemble selection methods," *Exp. Syst. Appl.*, vol. 106, pp. 141–153, Sep. 2018.

[27] K. W. De Bock, K. Coussement, and S. Lessmann, "Cost-sensitive business failure prediction when misclassification costs are uncertain: A heterogeneous ensemble selection approach," *Eur. J. Oper. Res.*, vol. 285, no. 2, pp. 612–630, Sep. 2020.

[28] H. R. Kadkhodaei, A. M. E. Moghadam, and M. Dehghan, "HBoost: A heterogeneous ensemble classifier based on the boosting method and entropy measurement," *Exp. Syst. Appl.*, vol. 157, Nov. 2020, Art. no. 113482.

[29] N. Ostvar and A. M. E. Moghadam, "HDEC: A heterogeneous dynamic ensemble classifier for binary datasets," *Comput. Intell. Neurosci.*, vol. 2020, pp. 1–11, Dec. 2020.

[30] P. Zyblewski, R. Sabourin, and M. Woźniak, "Preprocessed dynamic classifier ensemble selection for highly imbalanced drifted data streams," *Inf. Fusion*, vol. 66, pp. 138–154, Feb. 2021.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[32] R. M. Cruz, L. G. Hafemann, R. Sabourin, and G. D. Cavalcanti, "DESlib: A dynamic ensemble selection library in Python," *J. Mach. Learn. Res.*, vol. 21, no. 8, pp. 1–5, 2020.

[33] D. Dua and C. Graff. (2020). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

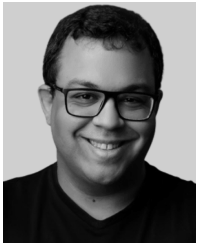
[34] T. T. Nguyen, A. V. Luong, M. T. Dang, A. W.-C. Liew, and J. McCall, "Ensemble selection based on classifier prediction confidence," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107104.

[35] A. Benavoli, G. Corani, and F. Mangili, "Should we really use post-hoc tests based on mean-ranks?" *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 152–161, 2016.

[36] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. 2006.

[37] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Mach. Learn.*, vol. 51, no. 2, p. 181, 2003.

[38] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 10, pp. 1863–1883, Oct. 2019.



PAULO R. G. CORDEIRO (Member, IEEE) was born in Recife, Pernambuco, Brazil, in 1987. He received the degree in computer engineering and the master's degree in computational intelligence from the University of Pernambuco (UPE), in 2012 and 2014, respectively. He is currently pursuing the Ph.D. degree in computational intelligence with the Federal University of Pernambuco, working with multiple classifier systems. He is also a Professor with Instituto Federal de Pernambuco (IFPE).

GEORGE D. C. CAVALCANTI (Senior Member, IEEE) received the Ph.D. degree in computer science from the Center for Informatics, Federal University of Pernambuco, Brazil, in 2005. He is currently a Full Professor with the Federal University of Pernambuco. He is also a Research Fellow with Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil. He is a member of the Pernambucana Academy of Science. His current research interests include machine learning, multiple classifier systems, meta-learning, and natural language processing.

RAFAEL M. O. CRUZ (Member, IEEE) received the Ph.D. degree in engineering from École de Technologie Supérieure (ÉTS), Montréal, QC, in 2016. He is currently an Associate Professor with ÉTS. His research interests include ensemble learning, data complexity, dynamic ensemble models, meta-learning, and biometrics.