IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 6, 2019, accepted October 18, 2019, date of publication October 23, 2019, date of current version November 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2949059

A Novel Diversity Measure and Classifier Selection Approach for Generating Ensemble Classifiers

MUHAMMAD ZOHAIB JAN, (Member, IEEE), AND BRIJESH VERMA
Centre for Intelligent Systems, School of Engineering and Technology, Central Queensland University, Brisbane, QLD 4000, Australia
Corresponding author: Muhammad Zohaib Jan (zohaibjan@gmail.com)

This work was supported by the Australian Research Council's Discovery Project Funding Scheme under Project DP160102369.

ABSTRACT Accuracy and diversity are considered to be the two deriving factors when it comes to generating an ensemble classifier. Focusing only on accuracy causes the ensemble classifier to suffer from "*diminishing returns*" and the ensemble accuracy tends to plateau; whereas focusing only on diversity causes the ensemble classifier to suffer in accuracy. Therefore, a balance must be maintained between the two for the ensemble classifier to achieve high classification accuracy. In this paper, we propose a novel diversity measure known as Misclassification Diversity (MD) and an Incremental Layered Classifier Selection (ILCS) approach to generate an ensemble classifier. The proposed approach ILCS-MD generates an ensemble classifier by incrementally selecting classifiers from the base classifier pool based on increasing accuracy and diversity. The benefits are in two folds 1) the generated ensemble classifier contains only those classifiers from the pool which can either maximize accuracy whilst maintaining or increasing the diversity, and 2) the generated ensemble classifier selects only a few classifiers from the base classifier pool thus reducing ensemble component size as well. The proposed approach is evaluated on 55 benchmark datasets taken from UCI and KEEL dataset repositories. The results are compared with five existing pairwise diversity measures, and existing state of the art ensemble classifier approaches. A significance test is also conducted to verify the significance of the results.

INDEX TERMS Ensemble classifiers, neural networks, multiple classifiers learning, diversity measures.

I. INTRODUCTION

Ensemble classifiers also known as "*multi-classifier systems*" are machine learning classification methods that are used to get better predictive performance over a single classifier. An ensemble classifier consists of multiple accurate and diverse base classifiers to form classification decisions which enables it outperform single classifiers [1]. Because of which ensemble classifiers are able to outperform single classifiers, moreover, a single classifier working well on one dataset might not work well on others, and this is known as the "*no free lunch*" theorem [2]. The main idea behind ensemble classifiers is that a committee of experts (classifiers) when suitably combined, that is they come up with a unanimous decision, the outcome is always better than a decision that is given by a single member (classifier). The classifiers within an ensemble must be diverse that is they must be making uncorrelated errors otherwise there is no point in combining classifiers with correlated errors [3].

Diversity and accuracy have been pointed out to be the two deriving factors when generating an ensemble classifier [4], [5]. Accuracy is the ability of a classifier to generate class labels as close to the ground truth as possible, and diversity is the difference between the classification abilities of various classifiers in the ensemble. A balance must be maintained between the two to generate an ensemble classifier that can perform well on unseen dataset. Although some authors argue that the main objective of any ensemble classifier is to achieve higher classification accuracy therefore, focus primarily should be on accuracy. However, others suggest that maximizing accuracy whilst maximizing diversity is a better strategy as eventually diverse ensembles perform better on noisy datasets [6]. Many different diversity measures have been proposed in [7]. Some strategies using

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

which diversity can be incorporated in an ensemble are as follows: (1) diversity creation by training classifiers on different input data sub samples for example in bagging, (2) diversity creation by training classifiers on different input features, for example in random forest, (3) diversity creation by using a set of structurally different classifiers that have different learning capabilities for example Support Vector Machine (SVM), Artificial Neural Networks (ANN), k-Nearest Neighbour (kNN), Decision Trees (DT), Naïve Bayes (NB) and Linear Discriminant Analysis (LDA).Some researchers have also utilized different optimization algorithms for example multi-objective optimization or genetic algorithms to find the best set of classifiers from the base classifier pool that can maximize both accuracy and diversity in an ensemble [8]–[12].

In this paper we generate an ensemble classifier by maximizing diversity and accuracy. We propose a novel pairwise diversity measure which computes diversity using the misclassification labels of two classifiers and a novel incremental layered classifier selection approach. The original contributions are as follow:

- A novel pairwise diversity measure is proposed. This diversity measure is used to determine whether a classifier should be selected from the pool to form the ensemble or not.
- A novel incremental classifier selection approach is proposed to generate an ensemble classifier using the proposed diversity measure.
- A comparative analysis of ensembles with different diversity measures is conducted.

The rest of the paper is organized as follows. Section II entails the current state-of-the-art ensemble classifier techniques. Section III discusses the proposed diversity measure and approach for generating an ensemble classifier. Section IV gives details about the datasets, experimental setup, experiments, results and analysis of results. Section V summarizes our findings and lays out future directions.

II. BACKGROUND

Ensemble classifiers have seen a lot of research in last two decades, primarily because ensemble classifiers are able to classify real world noisy datasets. Ensemble classifiers work better than single classifier models because they benefit from the ''*perturb and combine*'' strategy [13]. Some of the earliest milestones and hallmark work in ensemble classifiers were Bagging and Boosting [4], [14]. Bagging works by creating random subspaces from input data known as bags and trains base classifiers on different bags and combines them. Since base classifiers are trained on different data samples this incorporates diversity in the ensemble. Boosting works by successively training base classifiers on samples that are not classified correctly. Another state-of-the-art ensemble classifier proposed in [15] is Random Forest (RaF). RaF works by training base classifiers (DT) on random input features from the data and then suitably combines them. RaF has been very successful in classifying noisy real world datasets and variations of RaF over the years have been proposed by researchers. Ensemble classifier approaches can be divided into four main categories, 1) approaches that exploit the input sample space as in Bagging and Boosting, 2) approaches that exploit the feature space as in RaF, 3) hybrid approaches that exploit both feature and sample space, and 4) approaches that utilize different classifier combining strategies.

Besides bagging some authors have utilized clustering to generate a random subspace. Input data is used to generate sparse data clusters with unique and repeating records, these strategies are discussed in [16]–[21]. In [22] researchers proposed a progressive semi supervised ensemble learning to generate an ensemble classifier. In the proposed approach authors first generate a random subspace from input data then progressively enlarge the training set by incorporating an evolutionary sample selection process. In [23] authors employed a subspace and clustering methodologies to generate a subspace which has a balanced number of classes in different subspaces. Although many subspace learning strategies have been proposed in [24]–[26], very few approaches exist that utilize ensemble learning to maximize the final classification accuracy. Additionally, although the strategies discussed that have successfully utilized clustering to generate random subspace to train base classifiers, however, since datasets have randomness in them a clear distinction of how many clusters should be generated to create a diverse input space which in turn will generate an ensemble classifier that can achieve the highest classification accuracy is required.

For the second category of ensemble classifier approaches many variations of RaF are proposed in research. For example in [27] researchers proposed an Oblique DT with RaF to generate ensemble classifier. In a comprehensive benchmark study of 161 classifiers on UCI repository [28] datasets, it was concluded that a variation of RaF known as parallel RaF outperformed most of the classifiers. Similarly in another recent benchmark study of ensemble classifiers [29], it was concluded that Multi-Proximal RaF (MPRaF) which is a variation of RaF outperformed most of the ensemble classifiers and researchers suggested that it should be made the yard stick for future ensemble classifier comparisons. Rotation Forest (RoF) [14], [30] also uses DT to construct ensembles however RoF differs from RaF because it extracts features based on a rotation matrix and also all of the features are utilized in order to find the significant feature(s). More RaF based strategies are discussed in [31]–[33]. Although RaF approaches perform well on unseen datasets, the hyperplane constructed is piece wise orthogonal and the ensemble decision boundary turns out to be ''stage like'' [29]. Other strategies that can be employed to find the best set of features for generating ensemble classifiers are Principle Component Analysis (PCA) [34], LDA [35], and Neighbourhood Component Analysis (NCA) [36]. Many researchers generated ensemble classifiers by finding or selecting the significant or optimal set of features to generate ensemble

classifiers for further details readers can refer to [37]–[43]. Extracting features before classification can help in dimensionality reduction but that puts two constraints on the ensemble i) the overall performance of the ensemble depends on how well the feature selection has been performed, and ii) lower number of features before classification means less information for classifiers thus the ensemble might have weak learners.

The third category of ensemble approaches exploits both the feature space and the input space to generate ensemble classifiers. For example, in [44] authors proposed a hybrid ensemble approach and suggested that they are effective when dealing with small datasets with many features, and steps should be taken to augment the training set in order to train classifiers with sufficient learning base. Although application of hybrid ensemble approaches have been discussed with small datasets with large number of features however a different strategy needs to be adopted for large datasets with few features, because the input subspace is already sufficiently large and there is no requirement to augment it any further. Similarly, in [45] authors proposed a hybrid ensemble approach that utilizes stochastic search and clustering-based pruning to generate an ensemble. A multitude of base classifiers is trained with different parameters, so a pool of diverse classifiers is generated. Classifier clusters are then generated using the classification performance of classifiers. In this way classifiers performing similarly are clustered together. Finally, a single link clustering is applied to obtain the final partition of classifiers which are then utilized to generate the ensemble. Further hybrid ensemble learning approaches are discussed in [46]–[48].

The fourth category of ensemble approaches incorporates different strategies for classifier selection. For example in [49] researchers used genetic algorithms to generate an ensemble classifier for unbalanced datasets; in [50], [51] researchers used multi-objective Particle Swarm Optimization (PSO) to generate an ensemble classifier. In [52], [53] researchers used PSO as a model selection tool to select the best set of classifiers to generate an ensemble classifier. Further ensemble classifier approaches that incorporate multi-objective optimization, evolutionary algorithms, genetic algorithms, *etc.* are discussed in [54]–[62]. In [63] researchers proposed a clustering-based strategy to select a diverse set of classifiers from the pool of classifiers that is generated through bagging. In most of the discussed classifier selection based approaches the application of optimization is to select the best set of classifiers by optimizing for generalization performance; however if a rule based [64] machine learning method can be incorporated into the ensemble generation framework then the need for optimization can be eliminated. Therefore, in this paper a novel incremental layered classifier selection approach is proposed that utilizes the proposed diversity measure to generate an ensemble classifier. The proposed ensemble classifier exploits both the feature space and input subspace by selecting the best set of input features that can increase the learning capabilities of classifiers and generates a diverse and rich input subspace by clustering input data.

III. PROPOSED APPROACH

Diversity is considered an important factor when it comes to generating an ensemble classifier, as diverse classifiers when combined suitably together generate accurate ensembles. If we combine classifiers that have correlated errors then the ensemble classifier generated with n classifiers and an ensemble classifier generated with only 2 classifiers will have no difference. In order to achieve the benefit of having more than one classifier in an ensemble we must have diverse classifiers. Over the years many pairwise diversity measures have been proposed and many researchers have proposed different approaches to generate ensemble classifiers by balancing accuracy and diversity. In this paper we propose a pairwise diversity measure and use it in relation to clustering to generate the proposed ensemble classifier. The proposed approach is compared with different pairwise diversity measures to test the efficacy and other ensemble classifier approaches as well. We first discuss the proposed diversity measure in the next subsection.

A. PROPOSED DIVERSITY MEASURE

The proposed diversity measure is defined below. Let us define a dataset as $X = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where $x \in \mathbb{R}^d$ is a d-dimensional feature vector and $y \epsilon \{1, 2, \ldots, c$ is its corresponding class label having n number of samples and c classes.

Let i^{th} ensemble be denoted as $\varepsilon^i = \{^1, ^2, ^9$ and j^{th} ensemble be denoted as $\varepsilon^j = \{^1, ^2, ^9, ^{12}\}$, where ^{n} is a classifier from the pool. The purpose of introducing diversity measure here is to find how diverse two classifiers are and whether adding a new classifier in the ensemble causes any difference in the prediction capabilities of the ensemble; if it does then that classifier can be added to the ensemble otherwise adding it to the ensemble is not beneficial. In order to compute diversity firstly all classifiers from both the ensembles are utilized to classify the input feature vector of the datasets. Results are stored in two data matrices as shown below.

$$\varepsilon^i = \begin{bmatrix} y_1' c1 & y_1^{c2} & y_1^{c3} \\ y_1'^{c1} & y_2'^{c2} & y_2'^{c3} \\ \vdots & \vdots & \vdots \\ y_n'^{c1} & y_n'^{c2} & y_n'^{c3} \end{bmatrix} \tag{1}$$

$$\varepsilon^j = \begin{bmatrix} y_1'^{c1} & y_1'^{c2} & y_1'^{c9} & y_1''^{2} \\ y' c_2^1 & y' c_2^2 & y' c_2^9 & y' 2_2'^2 \\ \vdots & \vdots & \vdots & \vdots \\ y_n'^{c1} & y_n'^{2} & y_n'^{c9} & y_n'^{c2} \end{bmatrix} \tag{2}$$

Then a column wise mathematical mode of the data matrices is taken to get the predictions of each ensemble

(majority voting) as given below:

$$\varepsilon^i = \begin{bmatrix} y_1' \\ y_2 \\ \vdots \\ y_n' \end{bmatrix}, \quad \varepsilon^j = \begin{bmatrix} y'_1 \\ y'_2 \\ \vdots \\ y'_n \end{bmatrix} \tag{3}$$

After getting the final predictions of each ensemble, a column wise matrix of misclassified samples is generated for each ensemble. This matrix contains 1 for any misclassified label and 0 otherwise. For example, if dataset X has only 6 samples (for the sake of simplicity) and ensemble i and ensemble j misclassified the following labels $y_i^o =< y_2^o, y_3^o, y_6^o >$ and $y_j^o =< y_1^o, y_3^o, y_6^o >$ then their misclassification matrices can be written as follows:

$$y_i^o = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad y_j^o = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \tag{4}$$

The diversity MD^n is calculated using the following equation:

$$MD^n = \frac{\sum_{\forall i,j \in Z | i \neq j} I' \left\{ y_i^o \neq y_j^o \right\}}{\left| y_i^o \cup y_j^o \right|} \tag{5}$$

where y_i^o is the misclassified label of i^{th} ensemble ε^i on a validation dataset Z have x features and y class labels, the denominator $\left| y_i^o \cup y_j^o \right|$ gives the count of total number of errors caused by both the ensembles, and I' is an indicator function of misclassified labels between two ensembles given as:

$$I'\left(y_i^o, y_j^o\right) = \begin{cases} 0, y_i^o = y_j^o \\ 1, y_i^o \neq y_j^o \end{cases} \qquad \forall i, j \in landi \neq j \tag{6}$$

The output of the indicator function I for ensemble misclassification matrices in (4) can be written as follows:

$$I'\left(y_i^o, y_j^o\right) = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{7}$$

The diversity using (5) can be calculated as follows. The numerator $\sum_{\forall i,j \in Z | i \neq j} I' \left\{ y_i^o \neq y_j^o \right\}$ becomes 2 and denominator $\left| y_i^o \cup y_j^o \right|$ becomes 4. Therefore, the diversity D^n is 2/4 or 0.5, since there are only 2 different misclassified labels out of a total of 4. The diversity is 1 if the two ensembles in comparison made totally different errors and 0 otherwise. The proposed diversity helps in identifying those classifiers which bring new learning capabilities to the ensemble and helps in removing redundant classifiers. The steps involved to calculate the proposed diversity between two classifiers [1], and[1] are as follows:

Step1. Calculate the classification output for all feature vectors using classifiers [1] and [2].
Step2. Calculate misclassified labels of [1], and [2] by comparing classification output with ground truth and store result in matrix y_1 and y_2 respectively.
Step3. Take a logical XOR of both column matrixes y_1 and y_2; and store the new column matrix as I
Step4. Take sum of all elements of column matrix I to get a scaler quantity which is I'
Step5. Take union of y_1 and y_2, and compute mode.
Step6. Divide I by the mode of the union of y_1 and y_2

B. ENSEMBLE CLASSIFIER GENERATION FRAMEWORK

The proposed approach starts off by identifying the significant input features of the training data. Any insignificant feature(s) is discarded and the reduced feature set training data is utilized to generate a random subspace by generating data clusters incrementally. The number of clusters generated is a factor of the number of samples in the dataset. On all generated data clusters a set of structurally different and diverse base classifiers are trained. All trained classifiers are added to the base classifier pool and through a process of incremental classifier selection and an ensemble classifier is generated. A classifier is added to the ensemble in a layer if it is able to classify any misclassified input feature in the previous layer. The process of an incremental layered classifier selection is repeated until every single classifier in the base classifier pool has been compared and selected/discarded accordingly. Figure 1 shows the flowchart of the proposed approach and each component is described in the subsequent subsections.

1) SIGNIFICANT INPUT FEATURE SELECTION

The proposed approach first reduces the input feature space by selecting only the significant set of features that can maximize the overall classification accuracy of the ensemble. A NCA is performed on training dataset T to compute feature weights w of all input features and any feature having weight less than a relative loss is discarded.

The goal of NCA is to maximize the following objective function:

$$argmax(\sum_{i=1}^{n} P(y'_i|x_i)) \tag{8}$$

where $P(y'_i|x_i)$ is the probability of correctly classifying y given sample x.

The flowchart of the feature selection process is given in Figure 2.

2) RANDOM SUBSPACE GENERATION

Instead of generating a random subspace by creating random subsets of training data as in bagging the proposed approach generates a random subspace by creating data clusters that consists of unique data samples. A total of $D = K(K+1)/2$ data clusters are generated and $K = n$ where n is the number

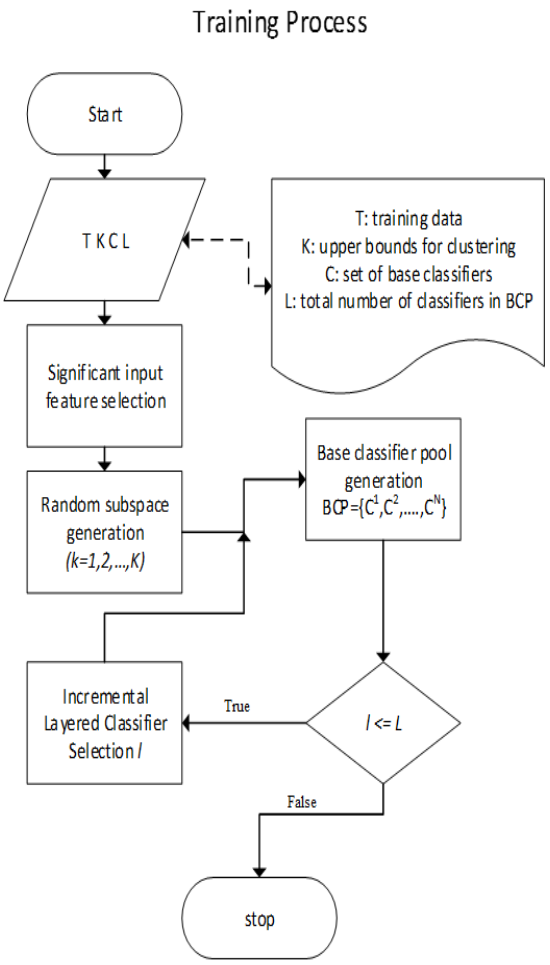

FIGURE 1. Flowchart of the proposed ensemble classifier generation framework.

of samples in the dataset. Generated clusters can be denoted as Ω_I and the set of total generated clusters is given as $\Omega = \{\Omega_1, \Omega_2, \ldots, \Omega_D\}$. Dataset can be partitioned into k clusters incrementally, where $k = 1, 2, ..K$. If in a particular iteration, k clusters are generated then clustering is achieved by minimizing the sum of the squared Euclidian distance of each observation with the cluster centroid and is given as:

$$argmin \sum_{i=1}^{n} \sum\nolimits_{\forall j \in k} \left(d\left(x_i, c_j\right)\right) \tag{9}$$

where x is a feature vector and c is a cluster centroid and $d\,(x, c)$ denotes the squared euclidean distance given as:

$$d\,(x, c) = (x - c)^2 \tag{10}$$

3) BASE CLASSIFIER POOL GENERATION

In the proposed approach a set of structurally different and diverse base classifiers $C = \{C^1, C^2, \ldots, C^N\}$ are trained on all generated data clusters (where $C^1 = ANN$ and $C^2 = SVM, etc.$). Since the total generated data clusters are D and for each data cluster a set of N classifiers is

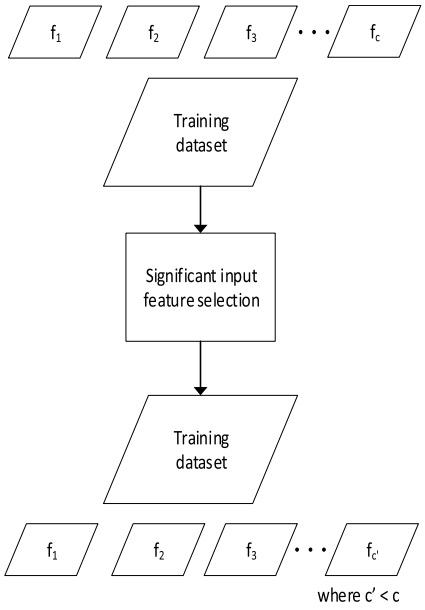

FIGURE 2. Significant input feature selection.

trained then after training on all generated data clusters the number of classifiers in the base classifier pool (BCP) is $Z = D \times N$. Each classifier has different learning capabilities as they use different architecture and learning algorithms and therefore introduces classifier diversity. Some classifiers are better trained than others as the data clusters have random set of records and depending on their learning capabilities classifiers are able to classify the unseen test set differently, therefore, classifiers should be selected from the BCP to generate an ensemble classifier that can achieve the highest classification accuracy on test set. The selection of best classifiers from the base classifier pool is described in the subsection to follow.

4) INCREMENTAL LAYERED CLASSIFIER SELECTION

In the proposed approach an ensemble classifier is generated by incrementally adding a classifier to the ensemble in each layer by means of the proposed diversity measure. Validation data set V is utilized and classification decisions of all classifiers from the BCP are obtained. In layer 1 a classifier is chosen randomly from the BCP and added to the ensemble. In layer 2 a second classifier from BCP is chosen and decisions of both classifiers from layer 1 and layer 2 are combined. Any decision fusion technique can be used but for simplicity we have used majority voting. Ensemble diversity is calculated using equation (5) and accuracy is computed as follows:

$$Acc^n = \sum_{i=1}^{n} I\left\{y_i^{'} = y_i\right\} \quad \forall y^{'}, y \in n \tag{11}$$

$$\text{where } I\left(y_i^{'}, y_i\right) = \begin{cases} 0, \&y_i^{'} \neq y_i \\ 1, y_i^{'} = y_i \end{cases} \tag{12}$$

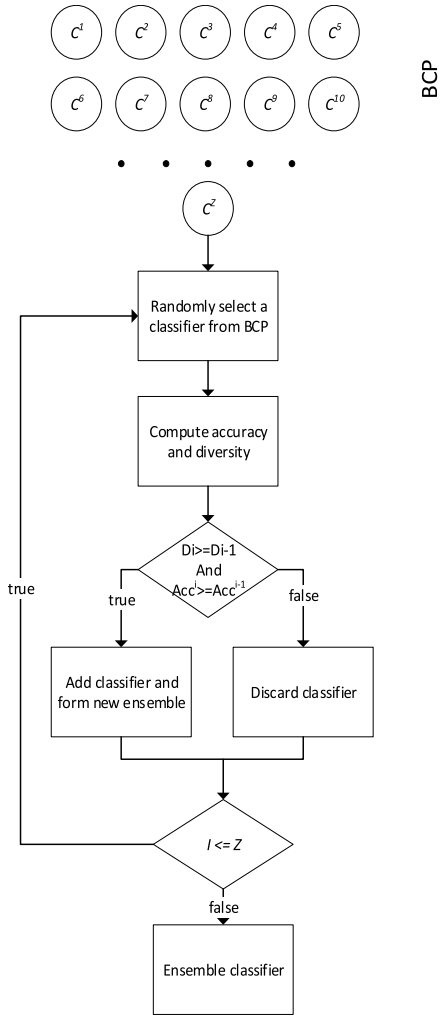

FIGURE 3. Incremental layered classifier selection.

and y is predicted class label of the ensemble obtained after combining decisions of all the classifiers in the ensemble ε^n and y is the actual class label.

If diversity D^i is higher than D^{i-1} and the ensemble ε^i has at least the same accuracy as ensemble ε^{i-1} then the ensemble ε^i is considered that has the new classifier from BCP in it and ensemble ε^{i-1} is discarded. If however D^i and D^{i+1} are same, then accuracy is compared and new ensemble is considered if the ensemble generated after adding the classifier achieves higher classification accuracy than the ensemble without adding the classifier, if not then the classifier is discarded. This process is repeated until every single classifier in the BCP has been compared. The process of ILCS is given in Figure 3 and steps for ILCS are given in Algorithm 1.

C. ENSEMBLE PREDICTION

To evaluate the performance of the generated ensemble classifier it is tested on the unseen dataset. The process of testing is given in Figure 4. First all of the classifiers selected in ILCS are utilized to classify the feature vector x of test set and

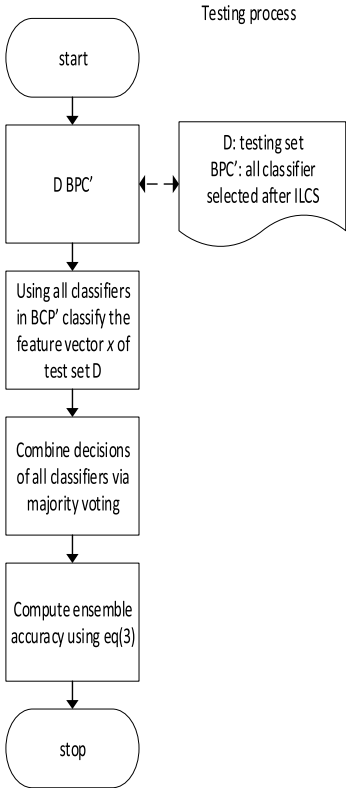

FIGURE 4. Testing process of ensemble classifier after ILCS.

all predictions y' are combined together via majority voting. The process of ensemble prediction is given in Figure 4

IV. EXPERIMENTATION

In this section we test the efficacy of the ensemble ILCS-MD that is generated using the proposed MD and ILCS. The 55 real world benchmark datasets taken from UCI [28] and KEEL [65] dataset repository are used for testing. ILCS-MD is also compared with five existing pairwise diversity measures given in [7] namely disagreement measure (DM), Q test (QT), double fault (DF), inverse correlation coefficient (IC), and interrater-k- (IK). The classification performance of the proposed approach is also compared with existing state-of-the-art ensemble approaches.

Table 1 provides summary of 55 real world datasets from KEEL and UCI repository. Number of samples, number of features, and number of class labels of each dataset is given in Table 1. It can be noted from Table 1 that a mix of datasets are chosen so that the proposed ensemble classifier can be tested thoroughly. Most of the datasets are challenging and are used in a large number of studies previously and therefore it enables us to compare the results of the proposed ensemble classifier with existing state-of-the-art ensemble classifier approaches.

In order to evaluate the quality of the proposed ensemble classifier we calculated classification accuracy on the predicted class labels of the unseen test set. The classification accuracy is calculated using equation (11). A 10-fold

TABLE 1. Summary of real world datasets from the UCI and KEEL dataset repository.

Datasets	# of samples	# of features	# of class labels	Datasets	# of samples	# of features	# of class labels
Adult	48842	14	2	Page Blocks	5473	10	5
Appendicitis	106	7	2	Parkinson	195	22	2
Australian	690	14	2	Pen digits	10992	16	10
Balance	625	4	3	Pima Diabetic	768	8	2
Banana	5300	2	2	Plan Relax	182	12	2
Bands	365	19	2	Seeds	210	7	3
Banknote	748	4	2	Segment	2310	19	7
Breast Cancer	683	9	2	Sonar	208	60	2
Bupa	345	6	2	Spam Base	460	57	2
Climate	540	18	2	Spect-f-heart	267	44	2
Contraceptive	1473	9	3	Stat log	2310	19	7
Dermatology	366	33	6	Stat image	6435	36	6
Diabetic Ret	1151	20	2	Teaching	151	5	3
DNA	3190	61	3	Thyroid	215	5	3
E.coli	336	7	2	Transfusion	748	4	2
Fertility	100	9	2	Two norm	7400	20	2
Glass	214	10	7	User knowledge	404	5	6
Haberman	306	3	2	Vehicle	946	18	4
Hayes-Roth	160	5	3	Vertebral 2C	310	6	2
Heart	270	13	2	Vertebral 3C	310	6	3
Hepatitis	80	19	2	Vowel	528	13	11
Ionosphere	351	33	2	Waveform V1	5000	21	3
Iris	150	4	3	Waveform V2	5000	21	3
Letter Recognition	20000	16	26	WDBC	569	30	2
Led7Digit	500	7	10	Wine	178	13	3
Liver	345	6	2	Yeast	1484	8	10
Mammographic	830	5	2	Zoo	101	17	7
Ozone Eight Hour	2534	72	2				

cross validation is adopted in experimentation to reduce the effect of randomness in the results. The proposed approach is implemented in Matlab [66] (version 2017 R1). Set of base classifiers ANN, SVM, kNN, DT, LDA, and NB with default parameters are used for training purposes. K-means clustering is used for data clustering with default parameters besides the following:

- Max iterations = 2400
- Distance measurement = *Squared euclidean*

For feature selection default implementation of NCA in Matlab was used with the following parameters:

- Solver = *stochastic gradient descent*
- Fit method = *exact*
- Standardization = *true*

A. COMPARISON WITH DIFFERENT PAIRWISE DIVERSITY MEASURE

We have compared the proposed MD with five other pairwise diversity measures that are QTest, DF, DM, IC, and IK. In order to compute the respective diversity measure the dissimilarity matrix is calculated which is given in Table 2.

TABLE 2. Dissimilarity matrix.

	D_j correct (1)	D_j wrong (0)
$D_i$correct (1)	N_{11}	N_{10}
$D_i$wrong (0)	N_{01}	N_{00}

Using the dissimilarity matrix the pairwise diversity measures are calculated as follows:

$$QTest\left(D_i, D_j\right) = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \tag{13}$$

$$DM\left(D_i, D_j\right) = \frac{N^{01} + N^{10}}{N^{11}+N^{10} + N^{01} + N^{00}} \tag{14}$$

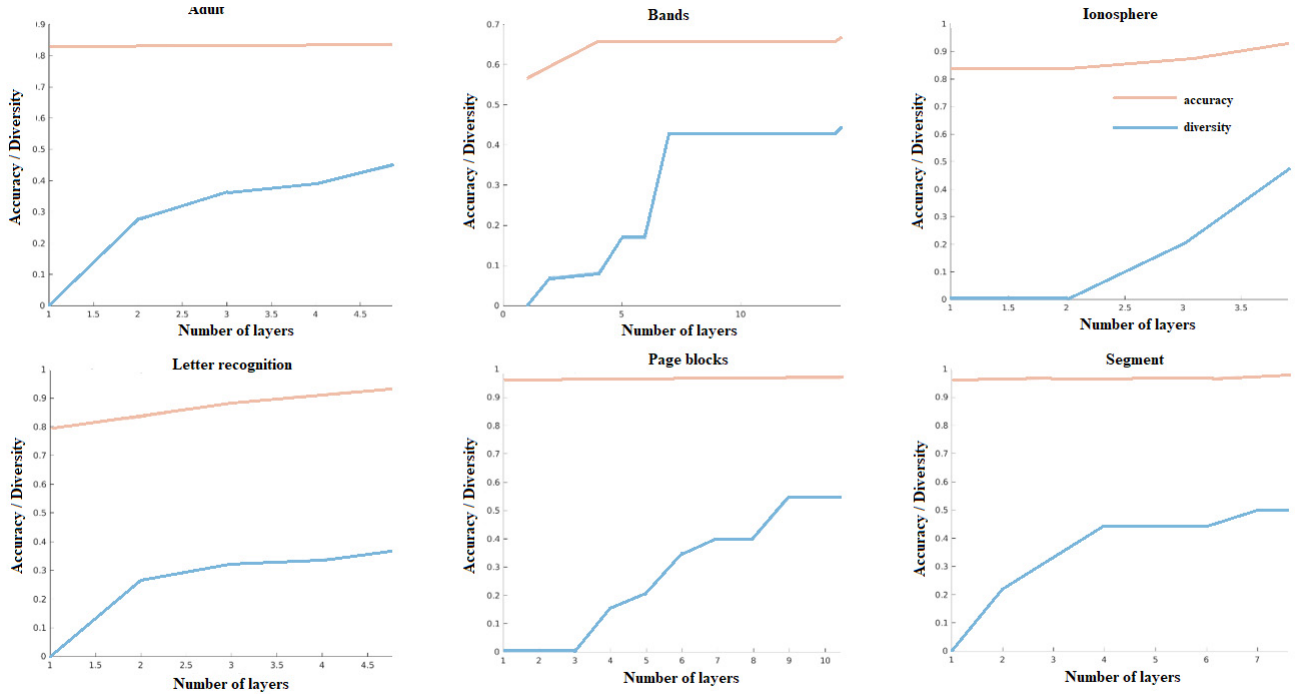

FIGURE 5. Effect of selecting classifiers on the basis of accuracy and proposed diversity measure in ILCS-MD.

Algorithm 1 Incremental Layered Classifier Selectio

Input: Base classifier pool BCP, Validation set V
Output: Ensemble classifier
Do:

1. Classify feature vector x of the validation set using all classifiers from the BCP.
2. Randomly select a classifier from BCP in layer 1.
3. Select another classifier from BCP, create ensemble ε^n and compute Acc^n and D^n
4. **if** $D^n \geq D^{n-1}$ and $Acc^n \geq Acc^{n-1}$ **then**
   a. Accept ensemble ε^n
   b. Remove classifier C^i from BCP
5. **else**
   a. Discard ensemble ε^n
   b. Remove classifier C^i from BCP
6. **Repeat** 3 to 4 $\forall \in Z$.

$$DF\left(D_i, D_j\right) = \frac{N^{00}}{N^{11}+N^{10}+N^{01}+N^{00}} \tag{15}$$

$$IK\left(D_i, D_j\right) = \frac{2 \times (N^{11}N^{00} - N^{10}N^{01})}{(N^{11}+N^{10})\left(N^{01}+N^{00}\right) + (N^{11}+N^{01})(N^{10}+N^{00})} \tag{16}$$

$$IC\left(D_i, D_j\right) = 1 - Cp\left(D_i, D_j\right) \tag{17}$$

TABLE 3. Average accuracy of ILCS on 55 benchmark datasets using MD, IC, IK, QT, DF, and DM diversity measures.

Proposed eensemble classifier with different diversity measures	*Average accuracy*
ILCS-MD	**89.33%**
ILCS-IC	73.75%
ILCS-IK	73.66%
ILCS-QT	73.64%
ILCS-DF	86.58%
ILCS-DM	88.10%

where $Cp\left(D_i, D_j\right)$ *iscalculatedasfollows* :

$$Cp\left(D_i, D_j\right) = \left(\frac{N^{11}N^{00} - N^{10}N^{01}}{\sqrt{(N^{11}+N^{10})\left(N^{01}+N^{00}\right)(N^{11}+N^{01})(N^{10}+N^{00})}}\right) \tag{18}$$

To calculate diversity between two ensemble classifiers the dissimilarity matrix is computed after classifying the feature vector of the training set. Using the predicted labels and the ground truth, the dissimilarity matrix values from Table 2 are calculated.

Average classification accuracies over 55 datasets achieved by ILCS-MD, ILCS-IC, ILCS-IK, ILCS-QT, ILCS-DF, and ILCS-DM are given in Table 3. It can be noted from Table 3 that ILCS-MD performed 21.12%, 21.27%,

TABLE 4. Training accuracy, training diversity, testing accuracy, and standard deviation of proposed ensemble approach on 55 UCI and KEEL repository datasets.

Datasets	Training Diversity	Training Accuracy	Test Accuracy	Std. Deviation	Datasets	Training Diversity	Training Accuracy	Test Accuracy	Std. Deviation
Adult	0.1764	0.8320	0.8379	0.0085	Page Blocks	0.2410	0.9710	0.9860	0.0039
Appendicitis	0.5000	0.5556	0.8955	0.0816	Parkinson	0.4706	0.4706	0.9074	0.0579
Australian	0.4597	0.4597	0.8261	0.0791	Pen digits	0.4929	0.4929	0.9970	0.0014
Balance	0.0000	0.9643	0.9647	0.0197	Pima diabetic	0.4058	0.4058	0.7656	0.0465
Banana	0.0790	0.8945	0.9040	0.0124	Plan Relax	0.0833	0.6563	0.7251	0.0257
Bands	0.1445	0.6406	0.7230	0.0855	Seeds	0.3750	0.9265	0.9750	0.0332
Banknote	0.5000	0.5000	0.9993	0.0023	Segment	0.3750	0.9638	0.9912	0.0037
Breast Cancer	0.1667	0.9731	0.9714	0.0213	Sonar	0.4167	0.8889	0.8752	0.0506
Bupa	0.2727	0.7204	0.7223	0.0626	Spam Base	0.4159	0.4215	0.8809	0.0427
Climate	0.4688	0.4688	0.9417	0.0394	Spect-f-heart	0.3913	0.4167	0.8085	0.0585
Contraceptive	0.1579	0.5152	0.6963	0.0240	Stat image	0.2567	0.7576	0.9697	0.0036
Dermatology	0.0000	0.9688	0.9897	0.0111	Stat log	0.5786	0.9819	0.9920	0.0016
Diabetic Retinopathy	0.4722	0.5388	0.7132	0.0475	Teaching	0.3750	0.7692	0.7042	0.0363
DNA	0.0680	0.8023	0.8836	0.0166	Thyroid	0.2838	0.9591	0.9817	0.0082
E.coli	0.2212	0.6833	0.9502	0.0165	Transfusion	0.4921	0.5224	0.7927	0.0381
Fertility	0.0000	0.8889	0.9000	0.0568	Two norm	0.4940	0.4955	0.9719	0.0258
Glass	0.3333	0.7895	0.9832	0.0184	User knowledge	0.1125	0.8819	0.9663	0.0188
Haberman	0.3929	0.7731	0.7578	0.0482	Vehicle	0.1111	0.8333	0.9013	0.0206
Hayes-Roth	0.4167	0.8095	0.8333	0.0635	Vertebral 2C	0.2000	0.7407	0.8839	0.0649
Heart	0.4375	0.8750	0.8481	0.0863	Vertebral 3C	0.1667	0.8889	0.9011	0.0252
Hepatitis	0.7333	0.8286	0.9000	0.0710	Vowel	0.4195	0.4326	0.9873	0.0108
Ionosphere	0.3333	0.8387	0.9232	0.0537	Waveform V1	0.4407	0.4444	0.9001	0.0248
Iris	0.0357	0.8516	0.9778	0.0314	Waveform V2	0.3647	0.4117	0.8983	0.0375
Letter Recognition	0.3864	0.4000	0.9400	0.0092	WDBC	0.0000	1.0000	0.9807	0.0149
Led7Digit	0.3713	0.8798	0.9965	0.0005	Wine	0.0000	1.0000	0.9963	0.0117
Liver	0.2444	0.6022	0.7242	0.0592	Yeast	0.0827	0.5990	0.9119	0.0109
Mammographic	0.4357	0.4662	0.8325	0.0298	Zoo	0.4444	0.7778	0.9867	0.0174
Ozone Eight Hour	0.2597	0.9639	0.9415	0.0084					

21.30%, 3.17%, and 1.39% better than ILCS-IC, ILCS-IK, ILCS-QT, ILCS-DF, and ILCS-DM respectively. DF and DM performed relatively better than other diversity measures with DM achieving the second highest average classification accuracy and DF the third, however, a significant performance boost was achieved in comparison to IC, IK and QT.

B. EFFECT OF USING ILCS – MD

Figure 5 shows the effect of incrementally selecting classifiers in layers in ILCS-MD. For the sake of simplicity 8 datasets were chosen on the basis of number of records, number of features, and number of classes. We can see from Figure 5 that in ILCS-MD there is a positive linear relation between accuracy and proposed diversity, also, the number of layers increment only if both accuracy and diversity increase or if one increases and other remains the same. In some cases, however, if both are same but adding a classifier enables the ensemble to classify a misclassified sample then that classifier is added therefore, causing an increment of a layer. The training accuracy, training diversity, testing accuracy, and standard deviation of the proposed approach on 55 datasets are given in Table 4.

C. COMPARISON WITH SINGLE CLASSIFIER MODELS

We have compared ILCS-MD with traditional single classifier approaches which include DISCR, SVM, kNN, DB, DT, and ANN. Default implementation of these approaches were used in Matlab with default parameters. For comparison 10-fold cross validation was conducted and average accuracies were calculated. The average accuracy over 55 datasets is given in Table 5. We can see that ILCS-MD achieves the highest average classification accuracy. ILCS-MD performs 4.59% better than DISCR, 3.42% better than SVM, 4.54% better than kNN, 8.27% better than NB, 5.36% better than DT, and 47.67% better than ANN.

TABLE 5. Average accuracy of the proposed approach in comparison with single classifier approaches.

Approach	Average Accuracy
ILCS-MD	**0.8621**
DISCR	0.8162
SVM	0.8279
kNN	0.8167
NB	0.7794
DT	0.8085
ANN	0.5838

D. COMPARISON WITH OTHER STATE OF THE ART ENSEMBLE CLASSIFIER APPROACHES

We have compared ILCS-MD with existing ensemble classifier approaches including RaF, boosting, WMV [67],

TABLE 6. Comparative analysis of ILCS-MD with MPRaF-T, highest accuracies given in bold.

Datasets	ILCS-MD	MPRaF-T	Datasets	ILCS-MD	MPRaF-T
Adult	**83.79**	83.40	Parkinson	90.74	**91.23**
Australian	82.61	**86.42**	Pima diabetic	**76.56**	75.91
Balance	**96.47**	89.02	Plan Relax	**72.51**	70.49
Banknote	**99.93**	99.91	Seeds	**97.50**	94.19
Breast Cancer	**97.14**	96.72	Sonar	**87.52**	82.12
Climate	**94.17**	92.06	Spam Base	88.09	**93.68**
DNA	88.36	**91.30**	Stat log	**99.20**	97.13
E.coli	**95.02**	85.46	Teaching	**70.42**	55.10
Fertility	**90.00**	87.90	Transfusion	**79.27**	78.05
Glass	**98.32**	94.21	Two norm	97.19	**97.47**
Haberman	**75.78**	72.39	User knowledge	**96.63**	91.20
Heart	**84.81**	83.74	Vehicle	**90.13**	76.30
Hepatitis	**90.00**	83.61	Vertebral 2C	88.39	**84.61**
Ionosphere	92.32	**92.68**	Vertebral 3C	**90.11**	83.61
Iris	**97.78**	97.60	Waveform V1	**90.01**	85.70
Mammographic	**83.25**	82.24	Waveform V2	**89.83**	85.44
Ozone Eight Hour	94.15	**97.12**	Wine	**99.63**	97.64
Page Blocks	**98.60**	97.28			

TABLE 7. Comparative analysis of ILCS-MD with WMV, highest classification accuracies given in bold.

Datasets	ILCS-MD	WMV	Datasets	ILCS-MD	WMV
Balance	**96.47**	82.90	Letter Recognition	**94.00**	90.90
Breast Cancer	**97.14**	95.80	Page Blocks	**98.60**	97.10
Bupa	**72.23**	68.80	Pen digits	**99.70**	97.50
Contraceptive	**69.63**	52.50	Pima diabetic	**76.56**	75.70
Dermatology	**98.97**	95.00	Segment	**99.12**	96.10
DNA	88.36	**94.10**	Sonar	**87.52**	75.90
E.coli	**95.02**	82.20	Stat image	**96.97**	89.50
Glass	**98.32**	71.00	Two norm	**97.19**	87.60
Heart	**84.81**	80.70	Vehicle	**90.13**	72.40
Hepatitis	**90.00**	81.00	Waveform V1	**90.01**	83.00
Ionosphere	**92.32**	91.50	Yeast	**91.19**	59.10
Iris	**97.78**	93.30	Zoo	**98.67**	91.20

PSEMISEL [22], and MPRaF-T [27]. For RaF, and boosting the default implementation in Matlab was used using the *"bag"* and *"LpBoost"* parameter for *fitcensemble* method; *"LpBoost"* was chosen for comparison because we have both multi-class and binary datasets in our experiments. As for other ensemble classifier approaches their results were taken directly from their respective papers. Due to different datasets used in experiments we have given comparative analysis in Table 6, Table 7 and Table 7. Of 35 common datasets ILCS-MD out performed MPRaF-T on 27 datasets. ILCS-MD performed on average 3.04% better than MPRaF-T.

Out of 24 datasets ILCS-MD out performed WMV on 23 datasets. On average ILCS-MD performed 9.78% better than WMV on 24 common benchmark datasets. The results are given in Table 7 with highest classification accuracies mentioned in bold. Lastly in comparison with

TABLE 8. Comparative analysis of ilcs-md with psemisel, highest classification accuracies given in bold.

Datasets	ILCS-MD	PSEMISEL
Banknote	**99.93**	88.73
Dermatology	**98.97**	91.49
Iris	**97.78**	89.61
Pima diabetic	**76.56**	63.11
Segment	**99.12**	92.26
Transfusion	**79.27**	59.81
Vehicle	**90.13**	61.79
Wine	**99.63**	93.66

TABLE 9. Comparative analysis of ILCS-MD with idafsen, highest classification accuracies given in bold.

Datasets	Test Accuracy	IDAFSEN
Balance	**0.9647**	0.9312
Breast Cancer	0.9714	**1.000**
Bupa	0.7223	**0.7888**
Hayes-Roth	**0.8333**	0.7956
Heart	0.8481	0.8481
Ionosphere	0.9232	**0.9428**
Pima diabetic	**0.7656**	0.7271
Seeds	0.9750	**1.000**
Spam Base	**0.8809**	0.8075
Wine	**0.9963**	0.8092

PSEMISEL, ILCS-MD performed 15.75% better and outperformed PSEMISEL on 8 common benchmark datasets.

TABLE 10. Comparative analysis of ILCS-MD with ebagts, highest classification accuracies given in bold.

Datasets	Test Accuracy	EBAGTS
Breast Cancer	0.9714	0.9635
Glass	0.9832	0.7508
Heart	0.8481	0.7911
Hepatitis	0.9	0.8352
Iris	0.9778	0.9622
Wine	0.9963	0.9631

We have also compared ILCS-MD with two pruning-based ensemble classifier approaches namely EBAGTS [68], and IDAFSEN [69]. The classification accuracies are taken directly from their respective papers and the results are given in Table 9, and Table 10. On average the proposed approach achieved a classification accuracy of 88.80% and IDAFSEN achieved a classification accuracy of 86.50%, achieving a performance gain of 2.3%. In comparison with EBAGTS, the proposed approach achieved an average of 94.61% and EBAGTS achieved an average of 87.77% resulting in the proposed approach achieving an average performance gains of 6.84%.

E. SIGNIFICANCE TESTING

In order to identify the significant difference among the results of different single classifier approaches, ILCS with

TABLE 11. Comparative analysis of ILCS-MD with Boosting, and Random Forest, highest classification accuracies given in bold.

Datasets	ILCS-MD	Boosting	Random Forest	Datasets	ILCS-MD	Boosting	Random Forest
Adult	83.79	85.03	**85.90**	Page Blocks	**98.60**	96.07	97.37
Appendicitis	**89.55**	84.45	87.73	Parkinson	90.74	**92.79**	90.29
Australian	82.61	82.90	**86.52**	Pen digits	**99.70**	96.32	99.16
Balance	**96.47**	82.08	83.98	Pima diabetic	**76.56**	70.46	74.99
Banana	**90.40**	85.21	88.92	Plan Relax	**72.51**	61.58	70.32
Bands	72.30	70.92	**74.54**	Seeds	**97.50**	86.00	93.97
Banknote	**99.93**	99.78	99.27	Segment	**99.12**	98.10	97.79
Breast Cancer	**97.14**	96.13	96.57	Sonar	**87.52**	84.57	85.52
Bupa	72.23	67.29	**73.59**	Spam Base	88.09	**94.26**	95.39
Climate	**94.17**	92.50	91.94	Spect-f-heart	80.85	80.57	**80.91**
Contraceptive	**69.63**	46.78	53.97	Stat image	**96.97**	83.37	91.72
Dermatology	**98.97**	95.79	97.50	Stat log	**99.20**	98.01	97.92
Diabetic retinopathy	**71.32**	63.94	66.81	Teaching	**70.42**	66.25	66.96
DNA	**88.36**	72.29	87.62	Thyroid	98.17	98.82	**99.67**
E.coli	**95.02**	83.93	86.91	Transfusion	**79.27**	65.64	75.40
Fertility	**90.00**	80.00	86.00	Two norm	**97.19**	96.05	97.11
Glass	**98.32**	94.45	97.71	User knowledge	**96.63**	92.80	91.80
Haberman	**75.78**	66.41	67.68	Vehicle	**90.13**	72.57	76.13
Hayes-Roth	83.33	**87.50**	81.25	Vertebral 2C	**88.39**	82.58	83.23
Heart	**84.81**	77.04	81.85	Vertebral 3C	**90.11**	84.52	82.90
Hepatitis	**90.00**	83.75	83.75	Vowel	**98.73**	84.95	96.57
Ionosphere	92.32	**94.31**	94.00	Waveform V1	**90.01**	81.76	84.70
Iris	**97.78**	97.33	94.67	Waveform V2	**89.83**	82.64	84.88
Letter recognition	**94.00**	69.40	71.80	WDBC	**98.07**	97.01	95.96
Led7Digit	**99.65**	90.13	96.68	Wine	**99.63**	82.97	97.22
Liver	72.42	66.40	**73.34**	Yeast	**91.19**	57.91	61.66
Mammographic	**83.25**	76.87	79.16	Zoo	**98.67**	90.55	97.00
Ozone Eight Hour	**94.15**	93.18	93.50				

TABLE 12. Wilcoxon signed rank test of proposed approach with single classifier approaches, other diversity measures and other ensemble approaches.

Approach	*p*-value
DISCR	3.3E-06
SVM	8.6E-06
kNN	1.0E-07
NB	3.3E-10
DT	3.4E-08
ANN	1.1E-10
WMV	8.1E-05
PSEMISEL	0.011
MPRaF-T	0.0005
EBAGTS	0.013
IDAFSEN	0.130
RaF	1.9E-06
Boosting	2.8E-08
ILCS-IC	2.63E-09
ILCS-IK	9.13E-09
ILCS-QT	7.74E-09
ILCS-DF	0.2631
ILCS-DM	0.0002

different diversity measures, and other ensemble classifier approaches we conducted non parametric tests [70] with a significance alpha value of 0.05 *i.e.* 95% significance. The results are given in Table 12. The tests are conducted to validate the results further and identify whether the alternate hypothesis *i.e.* the improvement in generalization performance is not by chance, can be accepted or not. The null hypothesis can be rejected with 95% confidence at *p*-values less than 0.05.

It can be noted from Table 12 that ILCS-MD performed significantly better than other classifier approaches besides ILCS-DF and IDAFSEN. Although the proposed approach performed 2.67% better than ILCS-DF and 2.3% better than IDAFSEN, however, the results are not statistically significant.

V. CONCLUSION

We introduced a novel pairwise diversity measure and a novel incremental layered classifier selection approach which selects classifiers in each layer based on the new diversity measure to generate an ensemble classifier. The propsed approach has been evaluated on 55 benchmark datasets from UCI and KEEL dataset respository. As shown in Table 3, the performance of ILCS with the proposed diversity measure is higher than other diversity measures, and, according to Table 6, 7, 8, and 9 the proposed ensemble has outperformed other state-of-the-art ensemble classifiers. A significance test has shown that results are statisticaly significant.

The results and analysis presented in this paper have shown that i) selecting classifiers from the base classifier pool on the basis of diversity and classification accuracy has a positive effect on the overall ensemble classification accuracy, ii) adding more classifiers to the ensemble classifier does not necessarily increase the performance of the ensemble, and a more suitable classifier selection process must be adopted, iii) diversity measure on the basis of misclassified labels works better than other pairwise diversity measures and iv) a robust classifier selection process has benefits in two folds; firstly, only those classifiers which can positively effect the ensemble classifier are selected, secondly, instead of having a very large ensemble component size, a small number of classifiers can achieve the same if not higher classification accuracy.

In future we will conduct further analysis of the effect of selecting classifiers on the basis of diversity and accuracy on reduction of ensemble component size, and ensemble classificaiton accuracy. We will also test the proposed approach on more real-world and benchmark datasets.

REFERENCES

[1] M. Wozniak, M. Grana, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," *Inf. Fusion*, vol. 16, pp. 3–17, Mar. 2014.

[2] D. H. Wolper and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, Apr. 1997.

[3] V. Cheplygina, D. M. J. Tax, and M. Loog, "Dissimilarity-based ensembles for multiple instance learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1379–1391, Jun. 2015.

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

[5] G. Rätsch, T. Onoda, and K.-R. Müller, "Soft margins for adaboost," *Mach. Learn.*, vol. 42, no. 3, pp. 287–320, 2001.

[6] M. Asafuddoula, B. Verma, and M. Zhang, "An incremental ensemble classifier learning by means of a rule-based accuracy and diversity comparison," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 1924–1931.

[7] 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, pp. 181–207, 2003.

[8] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *J. Mach. Learn. Res.*, vol. 12, pp. 2211–2268, Jul. 2011.

[9] G. Brown, J. L. Wyatt, and P. Tiňo, "Managing diversity in regression ensembles," *J. Mach. Learn. Res.*, vol. 6, pp. 1621–1650, Dec. 2005.

[10] C.-Y. Chiu, B. Verma, and M. Li, "Impact of variability in data on accuracy and diversity of neural network based ensemble classifiers," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Aug. 2013, pp. 1–5.

[11] C. Domingo and O. Watanabe, "MadaBoost: A modification of adaboost," in *Proc. Conf. Comput. Learn. Theory*, Jun. 2000, pp. 180–189.

[12] J. A. S. L. Filho, A. M. P. Canuto, and J. C. Xavier, "An analysis of diversity measures for the dynamic design of ensemble of classifiers," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2015, pp. 1–8.

[13] Y. Ren, L. Zhang, and P. N. Suganthan, "Ensemble classification and regression-recent developments, applications and future directions," *IEEE Comput. Intell. Mag.*, vol. 11, no. 1, pp. 41–53, Sep. 2016.

[14] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. 13th Int. Conf. Int. Conf. Mach. Learn.*, vol. 96, Jul. 1996, pp. 148–156.

[15] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.

[16] M. Asafuddoula, B. Verma, and M. Zhang, "A divide-and-conquer-based ensemble classifier learning by means of many-objective optimization," *IEEE Trans. Evol. Comput.*, vol. 22, no. 5, pp. 762–777, Oct. 2017.

[17] S. Fletcher and B. Verma, "Removing bias from diverse data clusters for ensemble classification," in *Proc. Int. Conf. Neural Inf. Process.*, 2017, pp. 140–149.

[18] Z. M. Jan, B. Verma, and S. Fletcher, "Optimizing clustering to promote data diversity when generating an ensemble classifier," in *Proc. Genetic Evol. Comput. Conf. Companion*, Jul. 2018, pp. 1402–1409.

[19] A. Rahman and B. Verma, "Novel layered clustering-based approach for generating ensemble of classifiers," *IEEE Trans. Neural Netw.*, vol. 22, no. 5, pp. 781–792, May 2011.

[20] A. Rahman and B. Verma, "Ensemble classifier generation using non-uniform layered clustering and genetic algorithm," *Knowl.-Based Syst.*, vol. 43, no. 2, pp. 30–42, May 2013.
[21] Z. Yu, X. Zhu, H.-S. Wong, J. You, J. Zhang, and G. Han, "Distribution-based cluster structure selection," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3554–3567, Nov. 2017.
[22] Z. Yu, Y. Lu, J. Zhang, J. You, H.-S. Wong, Y. Wang, and G. Han, "Progressive semisupervised learning of multiple classifiers," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 689–702, Feb. 2017.
[23] S. Nejatian, H. Parvin, and E. Faraji, "Using sub-sampling and ensemble clustering techniques to improve performance of imbalanced classification," *Neurocomputing*, vol. 276, pp. 55–66, Feb. 2018.
[24] H. Parvin, H. Alinejad-Rokny, B. Minaei-Bidgoli, and S. Parvin, "A new classifier ensemble methodology based on subspace learning," *J. Exp. Theor. Artif. Intell.*, vol. 25, no. 2, pp. 227–250, 2013.
[25] X. Xu, W. Li, D. Xu, and I. W. Tsang, "Co-labeling for multi-view weakly labeled learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1113–1125, Jun. 2016.
[26] M. Kim, "Greedy approaches to semi-supervised subspace learning," *Pattern Recognit.*, vol. 48, no. 4, pp. 1563–1570, Apr. 2015.
[27] L. Zhang and P. N. Suganthan, "Oblique decision tree ensemble via multi-surface proximal support vector machine," *IEEE Trans. Cybern.*, vol. 45, no. 10, pp. 2165–2176, Oct. 2015.
[28] K. Bache and M. Lichman. (2013). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml/
[29] L. Zhang and P. N. Suganthan, "Benchmarking ensemble classifiers with novel co-trained kernel ridge regression and random vector functional link ensembles," *IEEE Comput. Intell. Mag.*, vol. 12, no. 4, pp. 61–72, Nov. 2017.
[30] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A new classifier ensemble method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1619–1630, Oct. 2006.
[31] L. Zhang and P. N. Suganthan, "Random forests with ensemble of feature spaces," *Pattern Recognit.*, vol. 47, no. 10, pp. 3429–3437, Oct. 2014.
[32] B. H. Menze, B. M. Kelm, D. N. Splitthoff, U. Koethe, and F. A. Hamprecht, "On oblique random forests," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2011, pp. 453–469.
[33] Y. Ye, Q. Wu, J. Z. Huang, M. K. Ng, and X. Li, "Stratified sampling for feature subspace selection in random forests for high dimensional data," *Pattern Recognit.*, vol. 46, no. 3, pp. 769–787, 2013.
[34] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.
[35] D. Zhang, X.-Y. Jing, and J. Yang, "Linear discriminant analysis," in *Biometric Image Discrimination Technologies* (Biometric Image Discrimination Technologies). Philadelphia, PA, USA: IGI Global, 2006, pp. 41–64.
[36] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov, "Neighbourhood components analysis," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2005, pp. 513–520.
[37] R. Diao, F. Chao, T. Peng, N. Snooke, and Q. Shen, "Feature selection inspired classifier ensemble reduction," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1259–1268, Aug. 2014.
[38] K. Kim, H. Lin, J. Y. Choi, and K. Choi, "A design framework for hierarchical ensemble of multiple feature extractors and multiple classifiers," *Pattern Recognit.*, vol. 52, pp. 1–16, Apr. 2016.
[39] L. Nanni and A. Lumini, "Evolved feature weighting for random subspace classifier," *IEEE Trans. Neural Netw.*, vol. 19, no. 2, pp. 363–366, Feb. 2008.
[40] Y. Sun, S. Todorovic, and S. Goodison, "Local-learning-based feature selection for high-dimensional data analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1610–1626, Sep. 2010.
[41] X. Wang, J. Yang, and X. Teng, "Feature selection based on rough sets and particle swarm optimization," *Pattern Recognit. Lett.*, vol. 28, no. 4, pp. 459–471, 2007.
[42] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1656–1671, Dec. 2013.
[43] W. Yang, K. Wang, and W. Zuo, "Neighborhood component feature selection for high-dimensional data," *J. Comput.*, vol. 7, pp. 161–168, Jan. 2012.
[44] Z. Yu, L. Li, J. Liu, and G. Han, "Hybrid adaptive classifier ensemble," *IEEE Trans. Cybern.*, vol. 45, no. 2, pp. 177–190, Feb. 2015.
[45] A. Onan, S. Korukoğlu, and H. Bulut, "A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification," *Inf. Process. Manage.*, vol. 53, no. 4, pp. 814–833, Jul. 2017.
[46] H. Zhang, W. Liu, S. Wang, J. Shan, and Q. Liu, "Resample-based ensemble framework for drifting imbalanced data streams," *IEEE Access*, vol. 7, pp. 65103–65115, 2019.
[47] H. Zhang, W. Liu, J. Shan, and Q. Liu, "Online active learning paired ensemble for concept drift and class imbalance," *IEEE Access*, vol. 6, pp. 73815–73828, 2018.
[48] B. Verma and S. Z. Hassan, "Hybrid ensemble approach for classification," *Appl. Intell.*, vol. 34, no. 2, pp. 258–278, Apr. 2011.
[49] U. Bhowan, M. Johnston, M. Zhang, and X. Yao, "Evolving diverse ensembles using genetic programming for classification with unbalanced data," *IEEE Trans. Evol. Comput.*, vol. 17, no. 3, pp. 368–386, Jun. 2013.
[50] S. Z. Martínez and C. A. C. Coello, "A multi-objective particle swarm optimizer based on decomposition," in *Proc. 13th Annu. Conf. Genetic Evol. Comput.*, Jul. 2011, pp. 69–76.
[51] A. Peimankar, S. J. Weddell, T. Jalal, and A. C. Lapthorn, "Ensemble classifier selection using multi-objective PSO for fault diagnosis of power transformers," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jul. 2016, pp. 3622–3629.
[52] H. J. Escalante, M. Montes, and E. Sucar, "Ensemble particle swarm model selection," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2010, pp. 1–8.
[53] H. J. Escalante, M. Montes, and L. E. Sucar, "Particle swarm model selection," *J. Mach. Learn. Res.*, vol. 10, pp. 405–440, Feb. 2009.
[54] M.-J. Kim and D.-K. Kang, "Classifiers selection in ensembles using genetic algorithms for bankruptcy prediction," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 9308–9314, Aug. 2012.
[55] A. Chandra and X. Yao, "Ensemble learning using multi-objective evolutionary algorithms," *J. Math. Model. Algorithms*, vol. 5, no. 4, pp. 417–445, 2006.
[56] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*. Hoboken, NJ, USA: Wiley, 2001.
[57] S. Gu and Y. Jin, "Generating diverse and accurate classifier ensembles using multi-objective optimization," in *Proc. IEEE Symp. Comput. Intell. Multi-Criteria Decis.-Making (MCDM)*, Dec. 2014, pp. 9–15.
[58] A. Rosales-Pérez, S. Garcia, J. A. Gonzalez, C. A. C. Coello, and F. Herrera, "An evolutionary multiobjective model and instance selection for support vector machines with Pareto-based ensembles," *IEEE Trans. Evol. Comput.*, vol. 21, no. 6, pp. 863–877, Dec. 2017.
[59] C. Zhang, P. Lim, A. K. Qin, and K. C. Tan, "Multiobjective deep belief networks ensemble for remaining useful life estimation in prognostics," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2306–2318, Oct. 2017.
[60] K. Bhattacharjee, H. K. Singh, M. Ryan, and T. Ray, "Bridging the gap: Many-objective optimization and informed decision-making," *IEEE Trans. Evol. Comput.*, vol. 21, no. 5, pp. 813–820, Oct. 2017.
[61] M. F. Saeedian and H. Beigy, "Dynamic classifier selection using clustering for spam detection," in *Proc. IEEE Symp. Comput. Intell. Data Mining*, Mar./Apr. 2009, pp. 84–88.
[62] 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.
[63] H. Parvin, M. MirnabiBaboli, and H. Alinejad-Rokny, "Proposing a classifier ensemble framework based on classifier selection and decision tree," *Eng. Appl. Artif. Intell.*, vol. 37, pp. 34–42, Jan. 2015.
[64] S. M. Weiss and N. Indurkhya, "Rule-based machine learning methods for functional prediction," *J. Artif. Intell. Res.*, vol. 3, pp. 383–403, Dec. 1995.
[65] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera, "Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *J. Multiple-Valued Log. Soft Comput.*, vol. 17, pp. 255–287, Jun. 2011.
[66] MATLAB, *Statistics and Machine Learning Toolbox*. Natick, MA, USA: MathWorks Inc., 2013.
[67] L. I. Kuncheva and J. J. Rodriguez, "A weighted voting framework for classifiers ensembles," *Knowl. Inf. Syst.*, vol. 38, no. 2, pp. 259–275, Feb. 2014.
[68] H. Jamalinia, S. Khalouei, V. Rezaie, S. Nejatian, K. Bagheri-Fard, and H. Parvin, "Diverse classifier ensemble creation based on heuristic dataset modification," *J. Appl. Statist.*, vol. 45, no. 7, pp. 1209–1226, 2018.
[69] X. Zhu, Z. Ni, M. Cheng, F. Jin, J. Li, and G. Weckman, "Selective ensemble based on extreme learning machine and improved discrete artificial fish swarm algorithm for haze forecast," *Appl. Intell.*, vol. 48, no. 7, pp. 1757–1775, Jul. 2018.
[70] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945.

MUHAMMAD ZOHAIB JAN received the B.Sc. degree from Bahria University, in 2005, and the M.S. degree from the University of Sunderland, in 2007. He is currently pursuing the Ph.D. degree funded by the Australian Research Council (ARC) Discovery Project Grant. He is with the Centre for Intelligent Systems, Central Queensland University, Brisbane, Australia. He has published a number of articles in machine learning, ensemble classifiers, and data mining. His research interests include neural networks, evolutionary algorithms, machine learning, ensemble classifiers, and data mining. He is a Vice Chair of IEEE CIS Queensland Chapter.

BRIJESH VERMA is currently a Professor and the Director of the Centre for Intelligent Systems (CIS), School of Engineering and Technology (SET), Central Queensland University (CQU), Brisbane, Australia. He has authored/coauthored/co-edited 13 books, including *Roadside Video Data Analysis: Deep Learning*, nine book chapters and over 150 articles in areas such as neural networks, deep learning, evolutionary algorithms, pattern recognition, computer vision, image processing, digital mammography, and web information retrieval. His main research interests include computational intelligence and pattern recognition.

He is a member of the Australian Research Council (ARC) College of Experts. He was the Chair of the IEEE Computational Intelligence Society's Queensland Chapter and under his leadership the Chapter won Outstanding Chapter Award from IEEE CIS. He is an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS), an Editor-in-Chief of the *International Journal of Computational Intelligence and Applications* (IJCIA), and an Editorial Board Member of a number of international journals.