


<b>ITC 1/51</b> <b>Information Technology and Control</b> <b>Vol. 51/ No. 1/ 2022</b> <b>pp. 158-179</b> <b>DOI 10.5755/j01.itc.51.1.30083</b>	<b>Dual-Layer Deep Ensemble Techniques for Classifying Heart Disease</b>	
	Received 2021/11/05	Accepted after revision 2022/01/25
	 <a href="http://dx.doi.org/10.5755/j01.itc.51.1.30083">http://dx.doi.org/10.5755/j01.itc.51.1.30083</a>	

**HOW TO CITE:** Prakash, V. J., Karthikeyan, N. K. (2022). Dual-Layer Deep Ensemble Techniques for Classifying Heart Disease. *Information Technology and Control*, 51(1), 158-179. <https://doi.org/10.5755/j01.itc.51.1.30083>

# Dual-Layer Deep Ensemble Techniques for Classifying Heart Disease

## V. Jothi Prakash

Department of Information Technology; Karpagam College of Engineering; Coimbatore, Tamil Nadu, India;  
e-mail: jothiprakashv@gmail.com

## N. K. Karthikeyan

Department of Information Technology; Coimbatore Institute of Technology; Coimbatore, Tamil Nadu, India;  
e-mail: karthikeyan.nk@cit.edu.in

**Corresponding author:** jothiprakashv@gmail.com

The prevalence of heart disease is increasing at a rapid rate due to changes in food habits and lifestyle of people all over the world. Early prediction and diagnosis of this fatal disease is a highly excruciating task. Nowadays, the ensemble learning approaches are preferred owing to their effectiveness in performance when compared to the performance of a single classification algorithm. In this work, a Dual-Layer Stacking Ensemble (DLSE) technique and a Deep Heterogeneous Ensemble (DHE) technique to classify heart disease are proposed. The DLSE uses several heterogeneous classifiers to form an ensemble that is efficient as well as diverse. The proposed framework consists of two layers with the first layer consisting of three different base learning algorithms Naïve Bayes (NB), Decision Tree (DT), and Support Vector Machine (SVM). The second layer comprises of three different classifiers, Extremely Randomized Trees (ERT), Ada Boost Classifier (ABC) and Random Forest (RF). The second layer utilizes the results from the first layer to provide a diverse input for the three classifiers. Finally, the outcomes are fed to the meta-classifier Gradient Boosted Trees (GBT) to generate the final prediction. The DHE uses three deep learning models Convolutional Neural Networks with Bidirectional Long Short-Term Memory (CNN BiLSTM), Artificial Neural Network (ANN) and Recurrent Neural Network (RNN) with RF, ERT and GBT as the meta-learners. The performance of the proposed methods is compared with traditional state-of-the-art classifiers as well as existing ensemble learning and deep learning methods. The experimental outcomes show that the proposed DLSE and DHE methods perform exceptionally in terms of accuracy, precision and recall measures.

**KEYWORDS:** Deep Learning, Ensemble Techniques, Heart Disease, Machine Learning, Multiple Classifiers, Stacking Ensemble.

---

## 1. Introduction

The World Health Organization (WHO) has stated that nearly 31% of annual deaths occur because of heart disease [58]. The WHO has also estimated that more than 75% of those deaths occur in middle- and low-income countries [57]. This increase in heart disease is mainly based on the factors such as years of alcohol abuse, smoking, unhealthy food habits, stress, lack of physical activities etc. The changes in the environment such as increase in the level of air pollution, variations in the temperature also play a factor for prevalence of heart disease. It has been estimated that over 54 million people in India suffer from heart related ailments. The recent Coronavirus Disease 2019 (COVID-19) outbreak has raised concern over substantial increase in heart related ailments. The COVID-19 pandemic has increased the risk of severe infection in people with underlying heart disease or heart related problems. Therefore, there is a need for proper classification methodology not only for detecting heart disease but also for predicting the possibility of heart disease in future.

Machine learning [25] has been used extensively by researchers to classify and predict heart disease. Recent technology advancements in parallel processing [12], Graphical Processing Unit (GPU) technology [60] have urged many researchers to utilize this power to process the data more effectively. Ensemble methods [46] are always known to be highly effective in solving classification problems and are the most preferred techniques in the recent days. Ensemble techniques [17] rely on a collection of classifiers rather than focusing on the performance of a single classifier. These approaches build a meta-model based on the results of several diverse classifiers. This meta-model is then used to provide the final prediction outcome for the problem. A wide variety of machine learning algorithms have been developed over the recent years for solving classification and regression problems in real world. Most of the algorithms often deal with increasing the accuracy of classification and prediction.

Many researches were carried out in search of an algorithm that provides high accuracy. The ensemble approaches also fall into this category. Some of the ensemble techniques deal with model fusion, selection of the base learners dynamically, combination of same

or different base learners, bagging, applying voting scheme, stacked generalization among others. In this modern era, deep learning models have been successfully applied for classification and prediction tasks as they automate the process of feature extraction using the hierarchical feature learning approach.

---

## 2. Related Work

The ensemble approaches have proven to be more effective when compared to the performance of a single classifier. Some of the recent works in ensemble approaches are discussed in this section. Bashir et. al [7] discussed an ensemble approach using bagging for diagnosing heart disease. The approach used a multi-objective voting scheme for the final prediction result. Al-Barazanchi et. al [1] developed a bagging model for diagnosing neuromuscular disorders. The technique used a Decision Tree as the base learner and a voting mechanism was used to obtain the final prediction. Nilashi et. al [35] proposed an adaptive neuro-fuzzy ensemble model for predicting hepatitis disease. This model used a Self-Organizing Map (SOM) for clustering the data. The major drawback in this method is the computational time that is needed for diagnosing the disease.

Atallah and Al-Mousa [5] developed an ensemble method using the majority voting scheme. Four classifiers were used and the predictions were combined using hard voting method. This approach is just a combination of four basic classifiers using voting scheme and the performance was limited. Ani et. al [3] proposed a rotation forest-based ensemble technique for disease diagnosis. This technique used RF as the base learner. A two-tier classification ensemble for detecting coronary heart disease was explored by Tama et. al [53]. This technique used RF, Gradient Boosting Machine (GBM) and Extreme Gradient Boosting Machine (XGBoost) as separate homogeneous ensembles. Yekkala and Dixit [63] designed a Genetic Algorithm (GA) based ensemble for classifying heart disease. This technique used GA for selecting the attributes for classification. But this model was validated on only a single dataset. Brunese et. al [11] provided an ensemble learning method for detecting brain cancer. This method

used a weighted soft voting technique for generating the prediction.

A hybrid ensemble for detecting heart disease was designed by Zhenya and Zhang [67]. This ensemble used five heterogeneous classifiers and used Relief algorithm for dimensionality reduction. This method was tested using the statlog dataset from the UCI data repository. A swarm-based RF algorithm was contributed by Asadi et. al [4]. This technique used a multi-objective particle swarm optimization (MOPSO) combined with the RF algorithm for diagnosing heart disease. This research suggested the generation of diverse feature sets rather than the traditional bootstrapping of the samples.

An intelligent ensemble method for detecting coronary artery disease was contributed by Sapra et. al [48]. This approach focused on the cost-effectiveness and rapid prediction of heart disease. Marak et. al [31] proposed a semi-supervised ensemble for cancer diagnosis from gene expression data. This method combined the merits of semi-supervised learning and ensemble learning. The model was validated on eight gene expression datasets.

Baccouche et. al [6] proposed a deep learning ensemble model using Bidirectional Long Short-Term Memory (BiLSTM) and Bidirectional Gated Recurrent Unit (BiGRU) model with CNN for the prediction of heart disease. But this technique did not use the benchmark datasets to validate the proposed model. Ali et. al [2] proposed a deep learning-based ensemble model along with feature fusion for predicting heart disease. This approach used conditional probability and information gain for feature weight and feature elimination respectively.

Rath et. al [42] developed a deep learning method for predicting heart disease from the imbalanced ECG samples. This method used Generative Adversarial Network (GAN) model for dealing with the imbalanced samples and used an ensemble of LSTM and GAN for classification. Chen et. al [13] designed a Local Feature based LSTM (LF-LSTM) and a deep learning ensemble for detecting heart rate variability and acceleration. Plawiak et. al [38] proposed a deep ensemble method using genetic algorithm for cardiac arrhythmia detection using ECG signals. This method fused normalization, hamming window, cross-validation for constructing the layers of deep ensemble.

It can be seen from the related works that the ensemble learning can be either homogeneous or heterogeneous. The former will have a single base learning algorithm and the latter will have different base learning algorithms. The choice of the base learner is directly proportional to the effectiveness of the ensemble. This paves the way to carry out extensive research in the area of ensemble classification. Moreover, it can be seen that the deep ensemble models provide a higher performance by utilizing the merits of both ensemble and deep learning models. In this research, a dual layer stacking ensemble that uses three different base learning algorithms in each layer and a deep heterogeneous ensemble are proposed and are applied to diagnose heart disease.

---

## 3. The Proposed Ensemble Methodologies

### 3.1. Dual Layer Stacking Ensemble (DLSE)

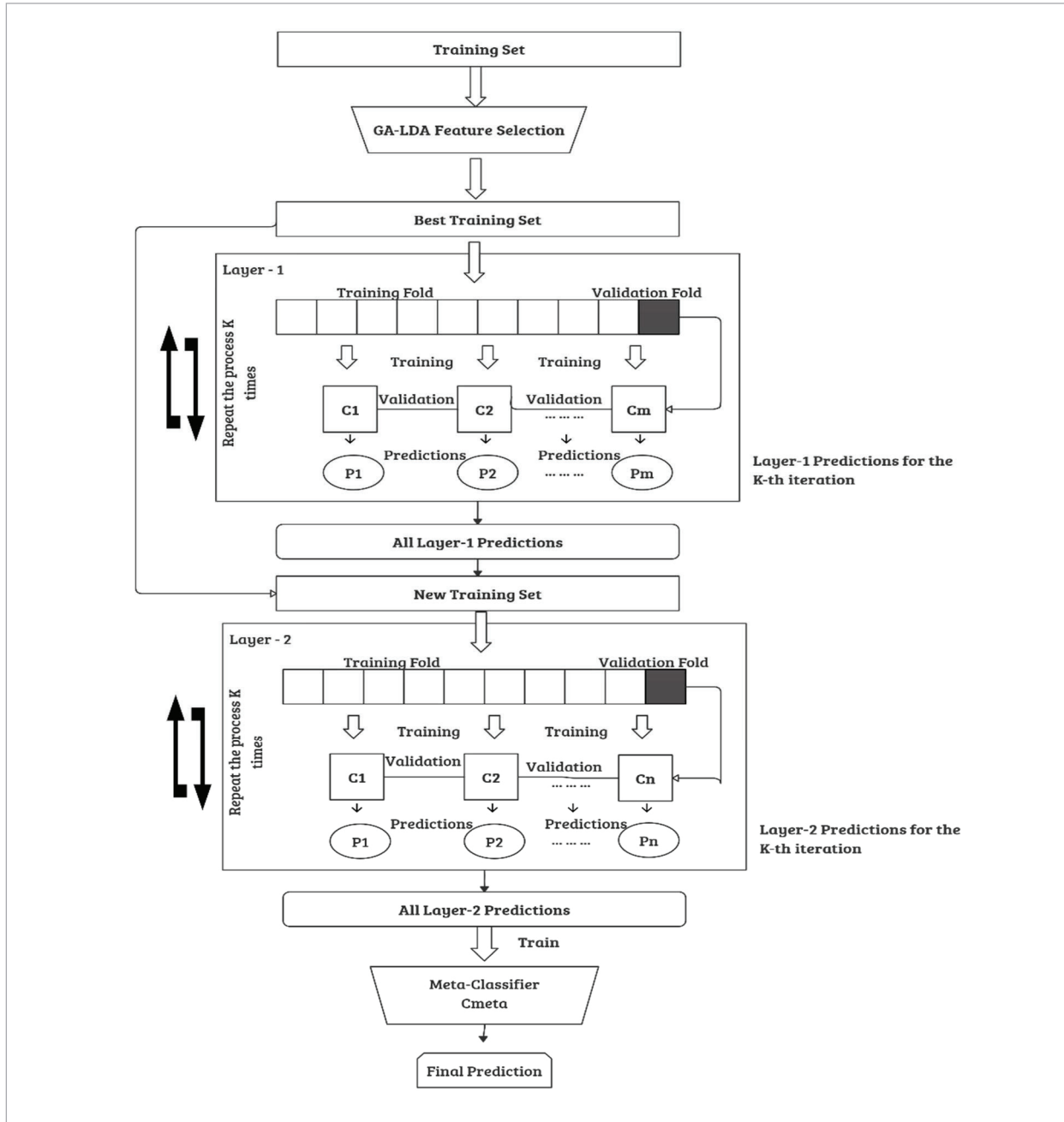
The proposed DLSE approach involves two layers of base learners and a final meta-learner to provide the final prediction. The Enhanced Evolutionary Feature Selection (EEFS) [40] algorithm is used to select the best feature set from the input training set. The best training set is then subjected to k-fold Cross Validation (CV) and is split into K disjoint subsets of equal size and one set from the K subsets is selected as the validation set. Once the K training sets are constructed the base learners in layer- 1 are trained and validated. We have used three classifiers NB, DT and SVM as the base learners in layer- 1. The prediction results of all the three classifiers are recorded and all the layer-1 predictions are then combined with the original training set and a new training set is given as input to layer-2 by combining the training set with the prediction matrix generated in layer-1.

Layer-1 can be considered as the feature generator for layer-2. This new training set is again subjected to k-fold CV and it results in K disjoint subsets of same size. Once again one subset is chosen at random as the validation set. Now the base learners in layer-2 are trained and validated. In layer-2 we have chosen ensemble classifiers ERT, ABC and RF as base learning algorithms. The second layer uses ensemble classifiers instead of traditional classifiers because the en-

semble-based classifiers always provide a better performance than the traditional classifiers [20, 22, 45, 47]. All layer-2 predictions are then used to train the

meta-classifier GBT. The meta-learner then provides the final prediction. The flow diagram of the proposed DLSE technique is shown in Figure 1.

**Figure 1**  
Flow diagram of the proposed DLSE Method



The pseudocode of the proposed Dual-Layer Stacking Ensemble (DLSE) method is shown in Algorithm 1 and the working principle is discussed. The dataset  $S = \{s_1, s_2, \dots, s_t\}$  along with the layer-1 base learners  $C_m$ , the layer-2 base learners  $C_n$ , and the meta-learner  $C_{meta}$  are provided as the input. Feature selection is applied on the dataset  $S$  to extract only the useful and important features. The feature selection of the data is performed by the GA-LDA [40] algorithm. The GA-LDA algorithm selects the best features  $S_{best}[s]$  from the given input dataset  $S$ . The data is then split into training set  $S_{train}[s]$  and testing set  $S_{test}[s]$  by applying the 80-20 rule. Then  $k$ -fold CV is applied.

We have chosen the value of  $K=10$  for validating the data. The CV yields  $K$  disjoint subsets  $S_{train}^k[s]$  with same size. The validation set  $S_{valid}[s]$  is chosen randomly from the training set  $S_{train}^k[s]$ . The training set  $S_{train}^k[s]$  is used for training the base learners  $C_m$  in layer-1. After training the validation set  $S_{valid}[s]$  is applied on the layer-1 base learners to obtain the prediction result  $S'_{v_i}[s]$  and a prediction matrix  $S'_v[s]$  is constructed by repeating the procedure  $K$  times. Then the results are fed to the second layer.

The second layer combines the training set  $S_{train}[s]$  generated in layer-1 with the prediction matrix  $S'_v[s]$ . This is done to ensure the learnings of layer-1 are propagated to layer-2. Again  $k$ -fold CV with  $K=10$  is applied to form the new training set. The CV yields another  $K$  disjoint subsets of same size  $S_{train}^k[s]$ . The validation set  $S_{valid}[s]$  is chosen randomly from the training set  $S_{train}^k[s]$ . The training set  $S_{train}^k[s]$  is used for training the base learners  $C_n$  in layer-2.

After training, the validation set  $S'_{valid}[s]$  is applied on the layer-2 base learners to obtain the prediction result  $S''_{v_i}[s]$  and this step is repeated  $k$  times to construct the prediction matrix  $S''_v[s]$ . The meta-learner  $C_{meta}$  is trained using the prediction matrix  $S''_v[s]$  constructed from layer-2. In the testing-phase, the test data  $S_{test}[s]$  is applied to each of the layer-1 and layer-2 base learners  $C_m$  and  $C_n$  and the prediction set  $S'_c[s]$  and  $S''_c[s]$  are constructed. Then the meta-learner is provided with the union of  $S'_c[s]$  and  $S''_c[s]$  to perform classification and the final prediction  $\hat{p}$  is returned.

### Algorithm 1

Pseudocode of proposed DLSE

---

#### Algorithm

---

**Input:**  $S = \{s_1, s_2, \dots, s_t\}, K=10, C_m, C_n, C_{meta}$

**Output:** Prediction result  $\hat{p}$

*Begin*

$S_{best}[s] =$  apply EEFS algorithm on  $S$  to select the best features.

Split  $S_{best}[s]$  into  $S_{train}[s], S_{test}[s]$

cross\_validation( $K, S_{train}[s]$ )  $\xrightarrow{\text{yields}}$   $S_{train}^k[s]$

$S_{valid}[s] \leftarrow S_{train}^k[s], r = \text{rand}(1, K)$

*//Layer-1*

**for** each  $i$  in  $C_m$  :

Train a base learner  $C_i$  on  $S_{train}^k[s]$

$S'_{v_i}[s] \leftarrow$  Apply  $C_i$  on  $S_{valid}[s]$

**end for**

Construct the prediction matrix

$S'_v[s] \leftarrow \{S'_{v_{k1}}[s], S'_{v_{k2}}[s], \dots, S'_{v_{km}}[s]\}$

*//Layer-2*

cross\_validation( $k, S_{train}[s] \cup S'_v[s]$ )  $\xrightarrow{\text{yields}}$   $S_{train}^{k'}[s]$

$S'_{valid}[s] \leftarrow S_{train}^{k'}[s], r = \text{rand}(1, k)$

**for** each  $i$  in  $C_n$  :

Train a base learner  $C_i$  on  $S_{train}^{k'}[s]$

$S''_{v_i}[s] \leftarrow$  Apply  $C_i$  on  $S'_{valid}[s]$

**end for**

Construct the prediction matrix

$S''_v[s] \leftarrow \{S''_{v_{k1}}[s], S''_{v_{k2}}[s], \dots, S''_{v_{km}}[s]\}$

*//Meta-Classifier*

Train  $C_{meta}$  based on  $(S''_v[s])$

*//Testing phase*

**for**  $i=1$  to  $m$  and  $j = 1$  to  $n$  do:

Apply  $S_{test}[s]$  on layer-1 base learners to obtain prediction set

$S'_c[s] \leftarrow \{C_1(s_i), C_2(s_i), \dots, C_m(s_i)\}$

Apply  $S_{test}[s] \cup S'_c[s]$  on layer-2 base learners to obtain prediction set

$S''_c[s] \leftarrow \{C_1(s'_i), C_2(s'_i), \dots, C_n(s'_i)\}$

**end for**

Apply  $C_{meta}$  to perform classification on

$S'_c[s] \cup S''_c[s]$

Return the final prediction  $\hat{p}_i$

*End*

---

### 3.1.1. Feature Selection Using EEFS

The dataset is feature selected using EEFS algorithm. EEFS is an evolutionary feature selection algorithm that utilizes the advantages of both GA and LDA. This algorithm treats each individual in a population as a binary string that encodes a feature subset. Therefore, for a dataset  $S$  of  $F$  features, it will be represented as an  $F$ -bit binary string. The '1' bits in the  $F$ -bit binary string correspond to the features that are selected and the '0' bits correspond to the features that are not selected. Table 1 shows the hyperparameters setting for EEFS algorithm.

**Table 1**

EEFS algorithm hyperparameters setting

Algorithm	Hyperparameters Setting
EEFS	population_size = 50 max_generations = 100 crossover probability = 0.8 mutation probability = 0.1 solver = 'svd'

The population size is set as 50 with maximum generations being assigned a value of 100. The crossover probability and mutation probability are set as 0.8 and 0.1 respectively. The selection scheme used is tournament with all the other parameters remaining in their default values. The solver for LDA is set as Singular Value Decomposition (SVD) and the remaining parameters are set with their default values.

### 3.1.2. Layer-1 Base Learners

The first layer of the proposed DLSE method consists of three simple classifiers. Three state-of-the-art classifiers NB [62], DT [27] and SVM [19, 64] have been used as base learning algorithms in layer-1. The Nave Bayes classification algorithm is well-known [29]. It estimates the conditional probability of each class given the observation and chooses the class with the highest posterior probability as the correct answer [50, 59]. It is employed in layer-1 because it requires the least amount of storage space to hold the probabilities in both the training and classification stages, making it a good fit for the high-dimensional datasets utilized in our research. SVM is based on statistical learning theory, which has since been improved by a number of other researchers. In SVM, kernel functions are

used to map training samples in high-dimensional space in a nonlinear way [56]. For mapping and optimizing the separation between data points, several kernel functions such as polynomial, Gaussian, and sigmoid are utilized. SVM's advantages, such as its success in high-dimensional spaces and flexibility in kernel function selection, have made it appealing for a variety of applications, including disease prediction, speech recognition, and text categorization. The DT classifier uses a tree-like graph and does not require any domain expertise. It creates conditional probabilities for research analysis and selects the optimal option for traversing from root to leaf, indicating distinct class separation [49]. It can be used in the medical industry to classify and forecast diseases. Moreover, the combination of NB, SVM and DT have proven to be very effective in classification [8, 9, 26]. Hence we have chosen the three classifiers for layer-1 of DLSE. The parameter setting for each of the algorithm is described in Table 2.

**Table 2**

DLSE Layer-1 base learner hyperparameters setting

Algorithm	Hyperparameters Setting
NB	-
DT	criterion = 'gain_ratio' max_depth = 10 min_split_size = 4 minimal_gain = 0.01
SVM	kernel='dot' max iterations =100000 convergence_epsilon = 0.001

The DT algorithm uses gain ratio as the criterion for selecting the attributes to split the tree and the maximum depth is set as 10 with the minimum split size being set as 4. The gain ratio  $G_{Ratio}$  measure is given by Equation (1),

$$G_{Ratio}(d_i, S) = \frac{I_{Gain}(d_i, S)}{H(d_i, S)}, \quad (1)$$

where,  $d_i$  is the attribute in training set  $S$ .  $H(d_i, S)$  is the entropy measure for the attribute  $d_i$  in the set  $S$ .  $I_{Gain}(d_i, S)$  is the information gain for the attribute  $d_i$  in the set  $S$  and is given by Equation (2),

$$I_{\text{Gain}}(d_i, S) = H(\hat{y}, S) - \sum_{v_{ij} \in \text{dom}(d_i)} \frac{|\sigma_{d_i=v_{ij}} S|}{|S|} * H(\hat{y}, \sigma_{d_i=v_{ij}} S), \quad (2)$$

where,  $\hat{y}$  is the target attribute,  $\frac{|\sigma_{d_i=v_{ij}} S|}{|S|}$  is the proportion of the number of elements in category  $i$  over the total number of records  $S$  and  $H(\hat{y}, S)$  is the entropy measure given by the Equation (3),

$$H(\hat{y}, S) = \sum_{c_j \in \text{dom}(\hat{y})} - \frac{|\sigma_{\hat{y}=c_j} S|}{|S|} \log_2 \frac{|\sigma_{\hat{y}=c_j} S|}{|S|} \quad (3)$$

The SVM uses dot kernel with a maximum iteration of 100000 along with the convergence epsilon value 0.001. The rest of the parameters of all the base learners remain in their default values.

### 3.1.3. Layer-2 Base Learners

In layer-2 three ensemble classifiers ERT [37], ABC [21] and RF [41] are used as base learning algorithms. Two of the most popular averaging methods are RF and ERTs. Before looking for the best features and split spots, it goes through two independent randomized algorithms. To begin, RF randomly selects a fixed number from the training set, similar to bagging [24]. Each decision tree is then grown using a randomly selected subset of input features. RF lowers variance and avoids overfitting by combining the two randomized techniques.

ERT is similar to RF. The bagging approach, on the other hand, is not employed when assigning training samples to each base learner. Instead, each base student is given the same set of training materials. Furthermore, the input feature and its splitting value are picked at random for the building of base learners, whereas RF looks for the highest discriminative thresholds. ABC allows predictors to be learned in a sequential manner. Iterative training is used to change weights for each observation and each base learner, lowering both variation and bias [15]. Moreover, the combination of these classifiers are proven to be effective [65] and hence we have chosen these three classifiers for layer-2 of DLSE. The parameter setting for the layer-2 base learners is shown in

Table 3. The ERT classifier uses a random subset just like RF but the random thresholds are set for each candidate feature and the best among the random thresholds is selected as the splitting criteria. The ERT uses averaging to minimize over-fitting and to maximize accuracy.

**Table 3**

DLSE Layer-2 base learner hyperparameters setting

Algorithm	Hyperparameters Setting
ERT	n_estimators = 200 criterion = 'gini' max_depth = 10 min_samples_split=2
ABC	base_estimator = 'Decision Tree' n_estimators = 200 learning_rate = 1
RF	n_estimators = 200 criterion = 'gini' max_depth = 10 min_samples_split=2

The ABC uses a decision tree as the base estimator with a learning rate of 1. All the learners are configured with 200 estimators and the other parameters remain with default values. The RF uses gini index as the criterion for split with a maximum depth of 10. The gini index  $G_{\text{Index}}$  is given by Equation (4),

$$G_{\text{Index}} = 1 - \sum_{g=1}^c D_g^2, \quad (4)$$

where,  $D_g$  is the proportion of samples that belongs to the class  $c$  for a particular tree node.

### 3.1.4. Meta-Learner

The GBT [34, 39] classifier is used as the meta-learner. The meta-learner is a regressor that allows optimization of least squares regression loss function  $L_c$ . At each stage of the regressor a regression tree is fit based on the negative gradient of the loss function  $L_c$ . The  $L_c$  is given by Equation (5), the negative gradient of the loss function  $L_c$ . The  $L_c$  is given by Equation (5),

$$L_c = \sum_{i=1}^n 1(p_i, E_{c-1}(e_i) + T(e_i)), \quad (5)$$

where,  $L_c$  is the loss for  $c^{th}$  ensemble,  $p_i$  is the prediction for input  $e_i$ ,  $E_{c-1}$  corresponds to the previous ensemble.  $T$  corresponds to the estimators used in the ensemble. A newly constructed tree  $T_c$  is fit accordingly to minimize the loss  $L_c$  given by previous ensemble  $E_{c-1}$  as shown in Equations (6)-(7).

$$T_c = \arg \min_T L_c . \tag{6}$$

By using Equation (5), we can rewrite Equation (6) as,

$$T_c = \arg \min_T \sum_{i=1}^n l(p_i, E_{c-1}(e_i) + T(e_i)). \tag{7}$$

**Table 4**

DLSE meta-learner hyperparameters setting

Algorithm	Hyperparameters Setting
GBT	n_estimators = 200 criterion = 'friedman_mse' max_depth = 10 learning_rate = 0.01

The parameter setting for the meta-learner is shown in Table 4. The number of estimators is set as 200 and the maximum depth is set to 10 with the learning rate of 0.01. The criterion for measuring the quality of the split used is the Friedman mean squared error  $R_{fmse}$  and is given by Equation (8),

$$R_{fmse} = \frac{n_1 * n_2}{n_1 + n_2} * (\bar{x}(1) - \bar{x}(2))^2, \tag{8}$$

where,  $n_1, n_2$  are the number of examples in each sub node and  $\bar{x}(n)$  corresponds to the mean output of the  $n^{th}$  sub node. The final prediction  $\hat{p}_i$  for the given input  $e_i$  is given by Equation (9),

$$\hat{p}_i = E_C(e_i) = \sum_{c=1}^C T_c(e_i), \tag{9}$$

where,  $C$  corresponds to the number of estimators  $n_{estimators}$  parameter and  $T_c$  are the estimators also called as weak learners. The meta-learner uses a fixed size of weak learners.

### 3.2. Deep Heterogeneous Ensemble (DHE)

The pseudocode of the proposed DHE algorithm is shown in Algorithm 2. The proposed DHE technique involves one layer of base learners and two layers of meta-learners to provide the final prediction. The first layer consists of three deep learning models CNN BiLSTM, ANN and RNN. The reason for selecting the base learners are deep learning models is from the fact that the deep learning models perform extremely well when the data and the feature sets are higher and also removes the need for manual feature extraction. The large dataset is split into training set  $U_{train}$  and testing set  $U_{test}$ . The training set  $U_{train}$  subjected to 10-fold CV to generate the  $K$  training sets  $U_{train}^k$ . One training set is chosen at random as the validation set  $U_{valid}^k$

**Algorithm 2**

Pseudocode of Deep Heterogeneous Ensemble (DHE)

#### Algorithm

**Input:**  $U = \{u_1, u_2, \dots, u_t\}, K=10, B_m, L1_{meta}, L2_{meta}$

**Output:** Prediction result  $\hat{f}$

*Begin*

Split the dataset  $U$  into  $U_{train}, U_{test}$

$cross\_validation(K, U_{train}) \xrightarrow{yields} U_{train}^k$

Randomly select a validation set  $U_{valid}^k$

**for** each  $j$  in  $B_m$ :

Train the base learner  $B_j$  on training set  $U_{train}^k$

Validate the trained base learner  $B_j$  using  $U_{valid}^k$

Record the predictions of  $B_j \xrightarrow{predictions} B_p$

**end for**

*// Level-1 Meta-Learners*

**for** each  $l$  in  $L1_{meta}$ :

Train the meta-learner  $L1_l$  based on the prediction matrix  $B_p$

Record the predictions of  $L1_l \xrightarrow{predictions} M_p$

**end for**

*// Level-2 Meta-Learner*

Train the meta-learner  $L2_{meta}$  based on the prediction matrix  $M_p$

Return the final prediction  $\hat{f}$

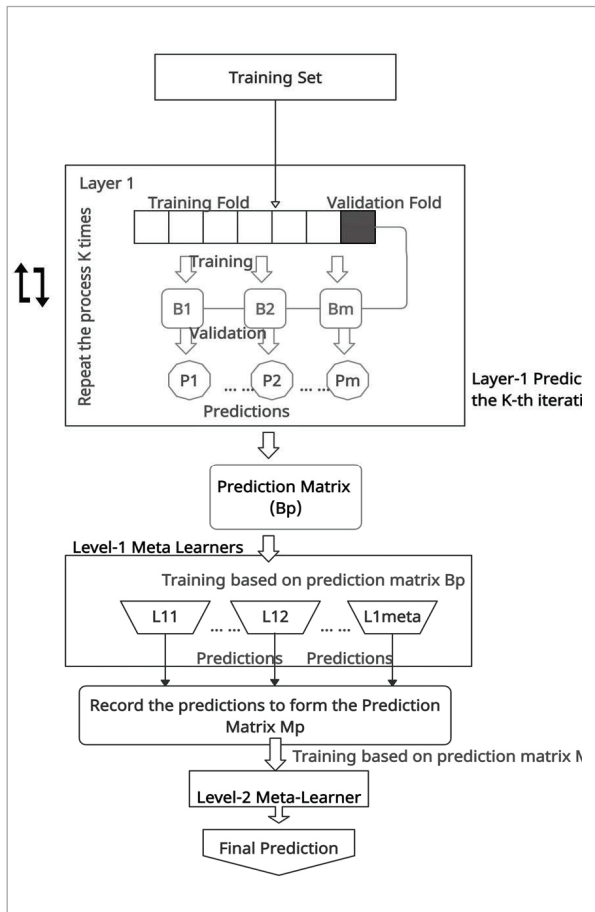
**End**



The base learners CNN BiLSTM, ANN and RNN are trained and validated using  $U_{train}^k$  and  $U_{valid}^k$  respectively. The predictions of each base learner is recorder to form the base learner prediction matrix  $B_p$ . This prediction matrix is then used to train the Level-1 meta learners of DHE. The RF and ERT algorithms are chosen as the level-1 meta learners. These two level-1 meta-learners are trained using the base learner prediction matrix  $B_p$ . Then the second level predictions are stored to form the level-1 meta learners prediction matrix  $M_p$ . This matrix is then fed as the input to the level-2 meta learner GBT. The level-2 meta learner is trained based on the predictions from Level-1 meta learner and the final prediction  $\hat{r}$  is returned as the output. The process flow of DHE is shown in Figure 2.

**Figure 2**

Flow diagram of the proposed DHE Method



### 3.2.1. Base Learners

The data is trained using three base learners CNN BiLSTM [43], ANN [55] and RNN [68]. The CNN BiLSTM is a hybrid bidirectional LSTM and CNN architecture. The CNN BiLSTM comprises of 8 convolutional layers, 4 dropout layers, 4 dense layers, 3 max pooling layers and 1 normalisation layer. The ANN consists of 4 dense layers, 3 dropout layers and 1 normalisation layer. Finally the RNN comprises of 3 dense layers, 2 dropout layers and 1 normalisation layer.

The proposed DHE method uses deep learning models as base learners and these base learners consist of a number of hyper parameters such as optimizer, learning rate, number of epochs and so on. Five hyper parameters are selected based on their effect on the performance of the deep learning models. The hyper parameters setting for all the base learners is shown in Table 5. In all the three models the activation function was selected as ReLU, the Rectified Linear Unit function. ReLU is one of the most widely used activation function which allows the deep learning models to be trained easily. The next important parameter is the number of epochs used to train the model. The epoch determine the number of times a training sample is selected in order to update the weights. This parameter will lead to over-fitting of the model on the training data set and hence needs to be optimised. The CNN BiLSTM model tend to be stable after 50 epochs and the ANN and RNN models were stable af-

**Table 5**

DHE base learner hyperparameters setting

Algorithm	Hyperparameters Setting
CNN BiLSTM	activation function = 'relu' dropout rate = 0.2 optimizer = 'Nadam' learning rate = 0.7 number of epochs = 50
ANN	activation function = 'relu' dropout rate = 0.2 optimizer = 'Nadam' learning rate = 0.7 number of epochs = 60
RNN	activation function = 'relu' dropout rate = 0.3 optimizer = 'Nadam' learning rate = 0.7 number of epochs = 60 recurrent dropout = 0.3

ter 60 epochs. Another parameter that helps to avoid over-fitting problem is the dropout rate. This parameter ensures the generalisation of the model. The dropout layer allows a fraction of input units to be dropped during training. It ranges between 0 and 1.

In all the three models the activation function was selected as ReLU, the Rectified Linear Unit function. ReLU is one of the most widely used activation function which allows the deep learning models to be trained easily. The next important parameter is the number of epochs used to train the model. The epoch determine the number of times a training sample is selected in order to update the weights. This parameter will lead to over-fitting of the model on the training data set and hence needs to be optimised. The CNN BiLSTM model tend to be stable after 50 epochs and the ANN and RNN models were stable after 60 epochs. Another parameter that helps to avoid over-fitting problem is the dropout rate. This parameter ensures the generalisation of the model. The dropout layer allows a fraction of input units to be dropped during training. The CNN BiLSTM model and ANN model showed highest performance for the dropout rate of 0.2 and the RNN model showed better performance for dropout rate 0.3. In order to reduce the loss function of the deep learning models an optimizer is used. All the three models performed extremely well for the optimizer 'Nadam' which is an Adam optimizer with Nesterov momentum. Finally the learning rate is another parameter that determines the optimization weights of the optimization algorithm. The learning rate for 'Nadam' optimization algorithm was varied and all the three deep learning algorithms showed stable performance for the learning rate of 0.7.

### 3.2.2. Level-1 and Level-2 Meta Learners

The level-1 meta learners used in DHE are RF and ERT. Both RF and ERT are tree based ensemble classifiers. The RF fits a several number of decision trees on different sub-samples of data. This method uses averaging to avoid over-fitting of data. The ERT works similar to the RF but uses random samples. The hyper parameter settings for both the meta learners is shown in Table 6.

The level-2 meta learner is a single meta estimator GBT. The GBT uses a regression tree based on a loss function shown in Equation (5). The parameter setting for GBT is shown in Table 7.

**Table 6**

DHE Level-1 Meta Learners hyperparameters setting

Algorithm	Hyperparameters Setting
ERT	n estimators = 300 criterion = 'gini' max depth = 10 min samples split = 5
RF	n estimators = 300 criterion = 'gini' max depth = 10 min samples split = 5

**Table 7**

DHE Level-2 Meta Learners hyperparameters setting

Algorithm	Hyperparameters Setting
GBT	n estimators = 300 criterion = 'friedman mse' max depth = 10 learning rate = 0.7

## 4. Performance Evaluation

The experiment is performed using a computer with Intel Core i7 processor having 16 gigabytes of Random-Access Memory (RAM) with a clock speed of 2.71 GHz and an NVIDIA GEFORCE RTX 2070 GPU. Five datasets are used to evaluate the proposed DLSE method out of which three datasets are from the University of California, Irvine data repository, the fourth dataset is from the ricco data repository and the last dataset is taken from the National Health and Nutrition Examination Survey (NHANES) repository. The datasets used to evaluate the proposed DLSE method are described in Table 8. Since the proposed DHE uses deep learning models it is evaluated using three larger datasets with more number of features and data samples. The three datasets used to evaluate the proposed DHE are MIT-BIH Arrhythmia Dataset, The PTB Diagnostic ECG Dataset and Longitudinal EHR dataset. The datasets used to evaluate the proposed DHE method are described in Table 9.

The performance of the model is evaluated using the traditional performance metrics precision, accuracy and recall. The efficiency of the proposed DLSE and DHE methods are measured using a confusion ma-

**Table 8**

Datasets used for evaluation of DLSE method

Dataset Name	No. of Instances	No. of Attributes	No. of Classes
Statlog Dataset [36]	270	14	2
SPECTF Dataset [30]	267	45	2
SPECT Dataset [14]	267	23	2
Eric Heart Dataset [44]	209	8	2
NHANES coronary heart disease Dataset [10]	37709	51	2

**Table 9**

Datasets used for evaluation of DHE method

Dataset Name	No. of Instances	No. of Attributes	No. of Classes
MIT-BIH Arrhythmia Dataset [32]	109446	188	5
PTB Diagnostic ECG Dataset [18]	14552	188	2
EHR dataset [66]	109490	89	2

trix. Here,  $T_n$  represents the True Negative,  $T_p$  corresponds to the True Positive,  $F_p$  and  $F_n$  represent the False Positive and False Negative values respectively. Based on these values the performance metrics are given by,

$$Accuracy = \frac{T_n + T_p}{T_n + T_p + F_n + F_p} \quad (10)$$

$$Precision = \frac{T_p}{T_p + F_p} \quad (11)$$

$$Recall = \frac{T_p}{T_p + F_n} \quad (12)$$

The proposed DLSE and DHE models are validated using k-fold cross validation. For this research the value of k is chosen as 10 making it 10-fold cross validation to estimate the performance of DLSE and DHE. The cross validation is applied on both the layers of DLSE.

#### 4.1. ANOVA Statistics

The statistical significance of the model is analysed by the ANalysis Of Variance (ANOVA) statistics. ANOVA Statistics is a statistical test that is used to determine the difference between group means and their variances, such as differences within and across groups. On the same data sets, the F -test is employed to measure the overall deviation pattern. The F-test results indicate which model best matches the supplied data set. The F-test, which is represented by the ANOVA F-test, is also used to determine whether the expected values of provided data sets differ from the values predicted by other classifiers. The value of F is roughly 1 if the null hypothesis is correct, but a large value of F causes the null hypothesis to be rejected. ANOVA condenses all of the data into a single number, F, and assigns a single p to the null hypothesis. The F-test statistics are calculated using the following formula:

$$F = \frac{\text{Between - Group Variability}}{\text{Within - Group Variability}} \quad (13)$$

The spread of a group of values/distribution is determined by its variability. There are two sorts of variability: between-group and within-group. The collaboration between the examples defines between-group variability, which is indicated by  $S^S(\text{BG})$  for sum of squares between groups. If the instances/samples have modest distances between them, the value of  $S^S(\text{BG})$  is small, and hence the grand mean is small. The differences within individual samples define within-group variability, which is expressed by  $S^S(\text{WG})$ , which is the sum of squares within groups. Because each sample is considered independently, there is no interaction between them. In the context of healthcare data, a within group indicates a single group of persons from many groupings. It can be a group of healthy people (class= 0) or patients with cardiac disease (class= 1). Thus, in this context, within group will indicate variability of attribute values within a group of heart disease patients or variability of attribute values within a group of healthy people. Between groups, on the other hand, depict multiple kinds of people from a same medical data collection. As an example, patients from both classes, those with and without heart disease, will be represented in the between group. In ANOVA statistics, the  $S^S$ ,  $d_p$ , MS, F,  $F_{\text{critical}}$  and p-value are determined. The sum

of squares ( $S^S$ ) is determined across groups using  $S^S(BG)$  variability and within groups using  $S^S(WG)$  variability using the formulas:

$$S^S(BG) = \sum n(x - \bar{X})^2 \quad (14)$$

$$S^S(WG) = \sum d_f * SD^2, \quad (15)$$

where  $x$  is the total values,  $\bar{X}$  is the mean of values,  $SD$  is the standard deviation, and  $n$  is one of many sample sizes. The variable  $d_f$  stands for "degree of freedom," which refers to the number of values in a data collection that are free to vary. Chi-square and hypothesis-testing statistics are widely employed with it. The degree(s) of freedom for the provided data set are used to determine the validity of the null hypothesis. Based on a number of variables and samples for the provided dataset, the degree of freedom can then be used to determine if a null hypothesis can be rejected. For both between-group and within-group comparisons, the  $df$  is calculated separately. The number of groups minus one equals the "between-group" degree of freedom, which is computed using the formula:

$$d_f = m - 1. \quad (16)$$

The number of groups is denoted by the letter  $m$ . The number of groups multiplied by the number of instances within each group, minus one, equals the degree of freedom "within-group." The following formula is used to compute it:

$$d_f = m(N - 1), \quad (17)$$

where  $N$  signifies the number of samples inside each group and  $m$  is the number of groupings.  $MS$  stands for mean square, and it is determined for the  $M^S(BG)$  group and the  $M^S(WG)$  group. By dividing the  $S^S(BG)$  by its degrees of freedom, the  $MS(B)$  is determined. By dividing the  $S^S(WG)$  by the degrees of freedom, the  $M^S(WG)$  is determined. The  $F_{critical}$  value is a function of the numerator degree of freedom, denominator degree of freedom, and significance level  $\alpha=0.05$ . The null hypothesis for ANOVA asserts that all groups have the same average value of the dependent variable (mean). It is always preferable to have an  $F$  value that is bigger than the  $F_{critical}$  value, since if this value is significant enough, we can reject the null hypothe-

sis in favor of the assumption that the classifiers we are comparing truly differ. ANOVA has long been a popular method for reviewing and interpreting medical data in the medical field. The importance of experimental data can also be determined using the p-value. The likelihood of finding a mean difference between groups given that the null hypothesis is true is defined as the p-value. A lower p-value, for example,  $p < 0.05$ , denotes a strong presumption against the null hypothesis and more significant results. For hypothesis tests, the p-value is particularly useful for weighing the strength of the evidence. A significant p-value suggests that there is insufficient evidence to reject the null hypothesis, which can never be rejected. The sample findings are usually observed at a significant level (threshold value), which is usually 0.05. However, the bayesian inference approach [16] suggests that this range of values may be optimistic, and thus establishes a new range in which  $p < 0.001$  denotes an algorithm's extreme significance level. By assuming that the null hypothesis is true, the p-value represents the chance of selecting a sample/value from a particular test dataset that is equal to or larger than observed test data sets. A p-value of 0.05 means that given the null hypothesis is true, there is only a 5% chance of drawing the sample being tested. The lower the p value, the more likely the null hypothesis will be rejected.

## 5. Results and Discussion

### 5.1. Evaluation of the Proposed DLSE Method

The proposed DLSE method is evaluated with traditional single classifiers and also with the existing ensemble techniques and the results are tabulated. We have also compared the DLSE method with a single layer ensemble method comprising of all the classifiers used in both layer-1 and layer-2 (NB, DT, SVM, LR, ERT, ABC and RF) of the proposed DLSE approach. In the proposed DLSE method, feature selection is applied on the dataset before applying the training set to layer-1. As mentioned before, the evolutionary feature selection algorithm EEFS is used for feature selection. The set of features selected using EEFS are shown in Table 10. The training set with selected features is then passed as input to the layer-1 of DLSE.

### 5.1.1. Evaluation with Single Classifiers

The performance of DLSE approach with single classifiers is shown in Table 11. It can be seen that for the Statlog dataset the accuracy of the proposed DLSE method is 94.21% which is the highest among all the other classifiers. Though the accuracy of NB, SVM and LR for the Statlog dataset is over 80%, the DLSE method performs better than these approaches. DT shows poor performance with accuracy of 75.19%. The precision and recall measure of DLSE for Statlog dataset is 95.21% and 96.08% respectively and are higher than all the other classifiers. The accuracy of NB, DT, SVM and LR for SPECTF dataset is 72.23%, 76.49%, 82.54% and 85.09% respectively. For this dataset also the proposed DLSE technique has achieved the highest accuracy of 92.34%. The proposed method also achieves highest precision value of 91.43% and recall value of 92.12% for the SPECTF dataset. The accuracy of the proposed method is 89.80% for SPECT dataset. NB obtains the lowest accuracy of 47.98% for the SPECT dataset. The precision

rate of the proposed method is 88.49% which is higher than the precision rates of NB (63.60%), DT (80.40%), SVM (69.38%) and LR (81.26%). The recall rate of the proposed technique is 81.99% for the SPECT dataset which is greater than the recall rates of NB (66.69%), DT (73.91%), SVM (78.92%) and LR (75.00%). The accuracy of NB, DT, SVM and LR for the Eric dataset is 78.02%, 76.57%, 78.98% and 78.98% respectively. For the Eric dataset also the proposed DLSE approach achieves the highest accuracy, precision and recall measures of 85.04%, 85.94% and 85.86% respectively. All the other approaches have precision and recall rates below 85%. The accuracy of the proposed method for NHANES dataset is 95.17%. This is the highest accuracy among the other approaches as the accuracy rate is almost 10% higher than the accuracy of NB (81.94%), DT (79.78%), SVM (85.80%) and LR (85.83%). The single classifiers have produced very poor precision and recall measures when compared to the proposed DLSE approach. The precision and recall rate of the proposed method is 89.66% and

**Table 11**

Evaluation of the proposed DLSE method with single classifiers

Dataset	Performance Metrics	Classification Techniques				
		NB	DT	SVM	LR	DLSE (proposed)
Statlog	Accuracy	82.96%	75.19%	82.96%	84.81%	94.21%
	Precision	83.56%	75.77%	84.81%	86.12%	95.21%
	Recall	82.58%	75.50%	81.92%	84.50%	96.08%
SPECTF	Accuracy	72.23%	76.49%	82.54%	85.09%	92.34%
	Precision	74.15%	70.51%	78.62%	82.57%	91.43%
	Recall	79.94%	67.17%	78.64%	81.50%	92.12%
SPECT	Accuracy	47.98%	85.00%	71.54%	85.74%	89.80%
	Precision	63.60%	80.40%	69.38%	81.26%	88.49%
	Recall	66.69%	73.91%	78.92%	75.00%	81.99%
Eric	Accuracy	78.02%	76.57%	78.98%	78.98%	85.04%
	Precision	78.52%	78.29%	81.02%	80.89%	85.94%
	Recall	77.10%	77.23%	77.51%	77.64%	85.86%
NHANES	Accuracy	81.94%	79.78%	85.80%	85.83%	95.17%
	Precision	58.66%	47.91%	47.91%	57.92%	89.66%
	Recall	63.07%	49.98%	49.99%	50.23%	85.43%

85.43% respectively. It can be seen that for all the datasets the proposed DLSE method outperforms all the single classifiers in terms of accuracy, precision and recall measures.

### 5.1.2. Evaluation with Other Ensemble Techniques

The results of the evaluation of the proposed DLSE method with the state-of-the-art ensemble techniques is shown in Table 12. We have compared the proposed method with Bagging ensemble with DT as the base learner, AdaBoost with DT as the base learner, RF and GBT methods. For the Statlog dataset, the accuracy of Bagging ensemble is 82.59%. AdaBoost and GBT both obtained an accuracy of 82.96% and RF achieved an accuracy of 80.74%. DLSE method achieved the highest accuracy of 94.21%. The precision and recall rates for Bagging ensemble is 83.28% and 82.25% respectively whereas for AdaBoost it is 84.00% and 82.67%, for RF it is 81.60% and 80.42% and for GBT it is 84.34% and 82.42%. DLSE obtained the highest precision rate of 95.21% and recall of 96.08%. The accuracy of the proposed DLSE method

is 92.34% for the SPECTF dataset. This is the highest accuracy when compared to Bagging (72.51%), AdaBoost (72.23%), RF (85.93%) and GBT (83.61%). Though precision rates of Bagging, AdaBoost, RF and GBT are 74.01%, 74.15%, 84.15% and 80.40% respectively the proposed method obtained the precision rate of 91.43% which is almost 17% higher than Bagging and AdaBoost, 7% higher than RF and 11% higher than GBT. The recall measure of DLSE is 92.12% and all the other ensembles obtained less than 80% recall rate. For the SPECT dataset, Bagging produced the lowest accuracy of 56.17% followed by AdaBoost with 71.99%, GBT with 83.85% and RF with 85.36%. The proposed DLSE method produced the highest accuracy of 89.80%. The precision and recall rate of DLSE is 88.49% and 81.99% which is the highest among all the other ensembles. AdaBoost produced the lowest precision and recall value of 44.05% and 53.18% respectively. DLSE obtained an accuracy of 85.04% for the Eric heart dataset. The other ensemble approaches achieved less than 80% accuracy. The precision and recall value of DLSE is 85.94% and 85.86% which is

**Table 12**

Evaluation of the proposed DLSE method with other ensemble techniques

Dataset	Performance Metrics	Classification Techniques				
		Bagging	AdaBoost	RF	GBT	DLSE (proposed)
Statlog	Accuracy	82.59%	82.96%	80.74%	82.96%	94.21%
	Precision	83.28%	84.00%	81.60%	84.34%	95.21%
	Recall	82.25%	82.67%	80.42%	82.42%	96.08%
SPECTF	Accuracy	72.51%	72.23%	85.93%	83.69%	92.34%
	Precision	74.01%	74.15%	84.15%	80.40%	91.43%
	Recall	79.84%	79.94%	79.75%	79.50%	92.12%
SPECT	Accuracy	56.17%	71.99%	85.36%	83.85%	89.80%
	Precision	63.60%	44.05%	79.88%	65.53%	88.49%
	Recall	66.69%	53.18%	74.16%	65.97%	81.99%
Eric	Accuracy	78.02%	77.55%	74.19%	78.02%	85.04%
	Precision	78.80%	78.43%	76.56%	80.92%	85.94%
	Recall	76.99%	76.96%	74.14%	76.48%	85.86%
NHANES	Accuracy	81.95%	82.93%	83.80%	82.68%	95.17%
	Precision	58.74%	57.37%	47.91%	56.83%	89.66%
	Recall	63.19%	57.86%	49.99%	57.86%	85.43%

again higher than all the other ensembles. Finally, for the NHANES dataset our proposed method achieved highest accuracy of 95.17% compared to Bagging (81.95%), AdaBoost (82.93%), RF (83.80%) and GBT (82.68%). The precision value of DLSE is 89.66% and all the other ensembles obtained a precision value less than 60%. The recall measure of DLSE is 85.43% which is again the highest value when compared to the rest of the ensembles as the recall value is 63.19% for Bagging, 57.86% for both AdaBoost and GBT and 49.99% for RF. In general, though the ensemble methods have better accuracy rates than that of the single classifiers, the proposed DLSE method outperforms them all in terms of accuracy, precision and recall.

### 5.1.3. Evaluation of Single-Layer and Dual-Layer Classification

We have also evaluated the proposed DLSE method with a single layered stacking ensemble using all the six base learners (NB, DT, SVM, LR, ERT, ABC and RF) with the meta-learner being GBT. The results are tabulated and are shown in Table 13. It can be seen that the DLSE method performs better than a single-layered ensemble of base learners in terms of all the performance metrics namely accuracy, precision and recall for all the datasets. The main advantage of using an ensemble of dual-layers is that it provides more flexibility than a single-layer ensemble. Since there are more

than one layer, we can use different classifiers in each layer resulting in a more refined classification. There is also a possibility for splitting an imbalanced classification problem in two relatively balanced problems. The dual-layer ensemble is also scalable for training and classifying hierarchically and can be applied to large medical datasets. The hierarchical classification always results in a better performance and quality classification than a simple flat structure. Moreover, the empirical evaluation shows that the dual-layered arrangement of classifiers outperforms the single-layered arrangement of classifiers.

### 5.2. Evaluation of the Proposed DHE Method

The proposed DHE method is evaluated against other popular ensemble techniques such as Boosting, Bagging, Stacking and the results are tabulated. It can be seen from Table 14 that the proposed DHE method outperforms the state-of-the-art ensemble techniques in terms of accuracy, precision and recall measures. The accuracy of the proposed DHE method for the MIT-BIH dataset is 99.50% which is the highest when compared to Bagging (92.31%), AdaBoost (88.48%) and Stacking (90.72%). The precision and recall measure for the DHE is 98.41% and 98.27% respectively which is also higher than the other ensemble models under consideration. The proposed DHE

**Table 13**

Comparison of the Single-Layer ensemble with DLSE

Dataset	Classification Techniques	Performance Metrics		
		Accuracy	Precision	Recall
Statlog	Stacking (Single-Layer)	81.62%	80.32%	79.59%
	DLSE (proposed)	94.21%	95.21%	96.08%
SPECTF	Stacking (Single-Layer)	87.53%	80.24%	77.28%
	DLSE (proposed)	92.34%	91.43%	92.12%
SPECT	Stacking (Single-Layer)	84.08%	71.34%	68.79%
	DLSE (proposed)	89.80%	88.49%	81.99%
Eric	Stacking (Single-Layer)	79.42%	80.56%	75.84%
	DLSE (proposed)	85.04%	85.94%	85.86%
NHANES	Stacking (Single-Layer)	89.88%	76.37%	77.86%
	DLSE (proposed)	95.17%	89.66%	85.43%

**Table 14**

Evaluation of the proposed DHE method with other ensemble techniques

Dataset	Performance Metrics	Classification Techniques			
		Bagging	AdaBoost	Stacking	DHE (proposed)
MIT-BIH Arrhythmia Dataset	Accuracy	92.31%	88.48%	90.72%	99.50%
	Precision	93.56%	82.31%	84.79%	98.41%
	Recall	90.15%	83.66%	82.37%	98.27%
PTB Diagnostic ECG Dataset	Accuracy	86.23%	86.57%	88.24%	99.87%
	Precision	84.17%	80.52%	82.76%	99.31%
	Recall	83.42%	82.92%	85.26%	99.01%
EHR Dataset	Accuracy	81.45%	84.79%	88.72%	98.03%
	Precision	79.92%	80.17%	83.63%	96.03%
	Recall	79.65%	79.18%	81.29%	96.13%

method achieves the highest accuracy of 99.87% for the PTB Diagnostic ECG dataset. The other ensemble techniques achieved accuracy rates below 90%. The precision value of DHE for the PTB dataset is 99.31% outlasting Bagging (84.17%), AdaBoost (80.52%) and Stacking (82.76%) by a very large margin. The recall measure is also 99.01% for the proposed DHE model which is higher than all the other ensemble models in comparison. Finally, for the EHR dataset the Bagging, AdaBoost and Stacking ensembles achieved an accuracy of 81.45%, 84.79% and 88.72% respectively. For this dataset also the proposed DHE achieved the highest accuracy of 98.03%. The precision and recall values of DHE for EHR dataset is 96.03% and 96.13% respectively which is again higher than that of Bagging (79.92%, 79.65%), AdaBoost (80.17%, 79.18%) and Stacking (83.63%, 81.29%).

### 5.3. Analysis of Statistical Significance

The statistical significance of the proposed models is discussed in this section. For a 95% confidence interval, we determined the  $p$ -value. The results show that the  $p$ -value is significantly lower than the selected threshold of 0.05. It also rejects the null hypothesis, implying that the proposed ensemble classifier outperforms competing classifiers across all datasets. The  $S^2$ ,  $d_p$ ,  $MS$  are determined and  $F$ ,  $F_{critical}$  and  $p$ -value are calculated and tabulated. Table 15 provide the findings of ANOVA statistics of DHE for MIT-BIH

Arrhythmia dataset, PTB Diagnostic ECG dataset and EHR Dataset. Table 16 provide the findings of ANOVA statistics of DLSE method for Statlog, SPECTF, SPECT, Eric and NHANES datasets. The suggested framework's results are statistically significant, according to ANOVA statistics. Table 15 and 16 present the ANOVA statistics of the proposed ensemble classifiers versus each individual classifier. Each individual classifier is compared to the proposed ensemble classifier, and "between-groups" and "within-groups" variables are calculated. The results show that  $F$  value is greater than  $F_{critical}$  for all classifiers, indicating that the proposed ensemble classifiers perform well. Furthermore, each classifier's  $p$ -value is less than 0.001, indicating that the results for heart disease prediction are strongly significant.

### 5.4. Evaluation of the Proposed Methods with Existing Approaches

The proposed DLSE method was evaluated against the existing ensemble approaches in the literature and the results are tabulated. Table 17 shows the comparison of accuracy of the proposed DLSE method with existing approaches. It can be seen that the proposed DLSE method obtained the highest accuracy for all the datasets used in the research. The proposed DHE method was evaluated against the existing deep ensemble techniques and the results are presented in Table 18. It can be seen that the proposed DHE meth-



**Table 15**

ANOVA statistics for the proposed DHE method against individual classifiers

ANOVA Statistics	Classification Techniques					
	CNN BiLSTM	ANN	RNN	ERT	RF	GBT
<b>MIT-BIH Arrhythmia Dataset</b>						
F	12.8025317	11.8100722	11.3633191	12.760754	10.9567079	11.4193088
F <sub>critical</sub>	3.8415013	3.8415013	3.8415013	3.8415013	3.8415013	3.8415013
p-value	<b>0.0003</b>	<b>0.0006</b>	<b>0.0007</b>	<b>0.0004</b>	<b>0.0009</b>	<b>0.0007</b>
<b>PTB Diagnostic ECG Dataset</b>						
F	15.05950311	15.1201337	11.186152	12.6743265	13.796277	13.3274772
F <sub>critical</sub>	3.841778376	3.84177838	3.8417784	3.84177838	3.8417784	3.84177838
p-value	<b>0.0001</b>	<b>0.0001</b>	<b>0.0008</b>	<b>0.0004</b>	<b>0.0002</b>	<b>0.0003</b>
<b>EHR Dataset</b>						
F	13.2963584	15.8097743	14.4670931	12.6601178	14.1753736	13.4516309
F <sub>critical</sub>	3.84150129	3.84150129	3.84150129	3.84150129	3.84150129	3.84150129
p-value	<b>0.0003</b>	<b>0.0001</b>	<b>0.0001</b>	<b>0.0004</b>	<b>0.0002</b>	<b>0.0002</b>

**Table 16**

ANOVA statistics for the proposed DLSE method against individual classifiers

ANOVA Statistics	Classification Techniques						
	NB	DT	SVM	ERT	ABC	RF	GBT
<b>Statlog Dataset</b>							
F	11.27924891	16.066406	16.397871	11.9736035	12.553568	12.4630172	13.362426
F <sub>critical</sub>	3.858801272	3.85880127	3.8588013	3.85880127	3.8588013	3.85880127	3.8588013
p-value	<b>0.0008</b>	<b>0.0001</b>	<b>0.0001</b>	<b>0.0006</b>	<b>0.0004</b>	<b>0.0005</b>	<b>0.0003</b>
<b>SPECTF Dataset</b>							
F	13.40314133	12.5224623	11.314488	15.4826697	12.232314	15.1067280	13.362426
F <sub>critical</sub>	3.858997525	3.85899752	3.8589975	3.85899752	3.8589975	3.85899752	3.8588013
p-value	<b>0.0003</b>	<b>0.0004</b>	<b>0.0008</b>	<b>0.0001</b>	<b>0.0005</b>	<b>0.0001</b>	<b>0.0003</b>
<b>SPECT Dataset</b>							
F	11.2707005	14.12799	12.2920171	14.8763209	11.90416	14.9268953	15.9066631
F <sub>critical</sub>	3.85899752	3.8589975	3.85899752	3.85899752	3.8589975	3.85899752	3.85899753
p-value	<b>0.0008</b>	<b>0.0002</b>	<b>0.0005</b>	<b>0.0001</b>	<b>0.0006</b>	<b>0.0001</b>	<b>0.0001</b>
<b>Eric Dataset</b>							
F	11.2248489	11.086864	11.4072125	11.367459	14.456157	11.3591665	11.2424261
F <sub>critical</sub>	3.86390928	3.8639093	3.86390928	3.86390928	3.8639093	3.86390928	3.86390928
p-value	<b>0.0009</b>	<b>0.0009</b>	<b>0.0008</b>	<b>0.0008</b>	<b>0.0002</b>	<b>0.0008</b>	<b>0.0009</b>
<b>NHANES Dataset</b>							
F	15.47247609	12.8866827	15.245849	14.4109847	14.181137	15.2955087	16.129520
F <sub>critical</sub>	3.841582128	3.84158213	3.8415821	3.84158213	3.8415821	3.84158213	3.8415821
p-value	<b>0.0001</b>	<b>0.0003</b>	<b>0.0001</b>	<b>0.0001</b>	<b>0.0002</b>	<b>0.0001</b>	<b>0.0001</b>

**Table 17**

Comparison of accuracy of the proposed DLSE method with existing approaches

S. No.	Classification Techniques	Accuracy				
		Statlog Dataset	SPECTF Dataset	SPECT Dataset	Eric Dataset	NHANES Dataset
1.	Heterogeneous ensemble [28]	85.36%	80.14%	79.24%	74.85%	82.94%
2.	Fog computing-based Ensemble [33]	86.45%	83.65%	82.58%	77.45%	81.47%
3.	Hybrid Recommender System [23]	88.74%	80.94%	83.34%	80.07%	86.47%
4.	Hybrid Ensemble [40]	93.65%	82.81%	84.95%	81.21%	83.32%
5.	DLSE (proposed)	<b>94.21%</b>	<b>92.34%</b>	<b>89.80%</b>	<b>85.04%</b>	<b>95.17%</b>

**Table 18**

Comparison of accuracy of the proposed DHE method with existing deep ensemble approaches

S. No.	Classification Techniques	Accuracy		
		MIT-BIH Arrhythmia Dataset	PTB Diagnostic ECG Dataset	EHR Dataset
1.	Ensemble of Neural Predictors [51]	91.36%	90.25%	91.23%
2.	Deep RNN [54]	93.62%	93.15%	92.05%
3.	Neural Networks Ensemble [61]	92.68%	94.36%	92.45%
4.	Optimal Stacked Ensemble [52]	95.32%	94.69%	95.84%
5.	DHE (proposed)	<b>99.50%</b>	<b>99.87%</b>	<b>98.03%</b>

od has obtained the highest accuracy for all the three datasets considered in this research. Overall, the results clearly portray the effectiveness of both DLSE and DHE methods in diagnosing heart disease.

## 6. Conclusion and Future Work

Ensemble techniques are in existence for over a decade and have been used in the domain of machine learning for classification and prediction. These approaches play a significant part in medical diagnosis for prediction and classification of diseases. In this work, dual-layer deep ensemble techniques namely, DLSE and DHE for heart disease classification and prediction were proposed. The proposed DLSE model was applied to five heart disease datasets and the results were analyzed. The proposed method was compared with both traditional single classifiers NB, DT, SVM and LR and also with state-of-the-art ensemble methods Bagging, AdaBoost, RF and GBT. The empir-

ical analysis shows that the proposed DLSE method excels in terms of accuracy, precision and recall. Also, the proposed DLSE was compared with a single-layer stacking ensemble comprising of all the machine learning approaches used in layer-1 and layer-2 of DLSE and the results further prove that the proposed dual-layered ensemble approach has higher accuracy than the traditional machine learning methods. The proposed DLSE method achieved the highest accuracy of 94.21% for Statlog dataset, 92.34% for SPECTF dataset, 89.80% for SPECT dataset and 85.04% for the Eric heart dataset. The highest overall accuracy achieved using DLSE method is 95.17% for the NHANES dataset. This strengthens the fact that hierarchical classification always results in a better performance and classification quality than a simple flat structure. The proposed DHE method was compared with other ensemble techniques Bagging, AdaBoost and Stacking. The performance evaluation shows that the proposed DHE method outperforms all the other ensemble methods by achieving an accuracy rate of

99.50% for the MIT-BIH Arrhythmia dataset, 99.87% for the PTB Diagnostic ECG dataset and 98.03% for the EHR dataset respectively. It can also be seen that the proposed DHE method is well-suited for larger datasets with a greater number of features. This also manifests the fact that the proposed DHE utilizes the merits of both deep learning and ensemble techniques. Moreover, at a 95% confidence interval, the F value and p-value derived from ANOVA statistics suggest that the results are statistically significant for all data sets. A major limitation of the proposed approaches is the time taken for training. The training time was not taken into account in the experiment.

Ensemble classifiers require more training time than individual classifiers. Overall, when compared to individual classifiers and earlier research, the suggested ensemble achieved much superior results, suggesting that it may be employed as a viable alternative tool in medical decision-making for heart disease detection.

In future, the proposed DLSE and DHE methods can be applied in classification and prediction of different diseases such as cancer and diabetes. Measures to reduce the training time of DLSE and DHE by applying parallel processing can be investigated. Furthermore, increasing the number of layers in the proposed method and analyzing the performance can also be explored.

## References

1. Al-Barazanchi, K. K., Al-Neami, A. Q., Al-Timemy, A. H. Ensemble of Bagged Tree Classifier for the Diagnosis of Neuromuscular Disorders. In Fourth International Conference on Advances in Biomedical Engineering (ICABME), 2017, 1-4. <https://doi.org/10.1109/ICABME.2017.8167564>
2. Ali, F., El-Sappagh, S., Islam, S. R., Kwak, D., Ali, A., Imran, M., Kwak, K.S.A. Smart Healthcare Monitoring System for Heart Disease Prediction Based on Ensemble Deep Learning and Feature Fusion. *Information Fusion*, 2020, 63, 208-222. <https://doi.org/10.1016/j.inffus.2020.06.008>
3. Ani, R., Jose, J., Wilson, M. and Deepa, O. S. Modified Rotation Forest Ensemble Classifier for Medical Diagnosis in Decision Support Systems, Springer, 2018. [https://doi.org/10.1007/978-981-10-6875-1\\_14](https://doi.org/10.1007/978-981-10-6875-1_14)
4. Asadi, S., Roshan, S., Kattan, M.W. Random Forest Swarm Optimization-Based for Heart Diseases Diagnosis. *Journal of Biomedical Informatics*, 2021. <https://doi.org/10.1016/j.jbi.2021.103690>
5. Atallah, R., Al-Mousa, A. Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method. In 2nd International Conference on New Trends in Computing Sciences (ICTCS), 2019. <https://doi.org/10.1109/ICTCS.2019.8923053>
6. Baccouche, A., Garcia-Zapirain, B., Castillo Olea, C., Elmaghraby, A. Ensemble Deep Learning Models for Heart Disease Classification: A Case Study from Mexico, *Information*, 2020, 11(4), 207. <https://doi.org/10.3390/info11040207>
7. Bashir, S., Qamar, U., Khan, F. BagMOOV: A Novel Ensemble for Heart Disease Prediction Bootstrap Aggregation with Multi-Objective Optimized Voting. *Australasian Physical & Engineering Sciences in Medicine*, 2015, 38(2), 305-323. <https://doi.org/10.1007/s13246-015-0337-6>
8. Bashir, S., Qamar, U., Khan, F.H. A Multicriteria Weighted Vote-Based Classifier Ensemble for Heart Disease Prediction. *Computational Intelligence*, 2015. <https://doi.org/10.1111/coin.12070>
9. Bashir, S., Qamar, U., Younus Javed, M. An Ensemble Based Decision Support Framework for Intelligent Heart Disease Diagnosis. *International Conference on Information Society*, 2014. <https://doi.org/10.1109/i-Society.2014.7009056>
10. Beck, J. D., Moss, K. L., Morelli, T. and Offenbacher, S. Periodontal Profile Class Is Associated with Prevalent Diabetes, Coronary Heart Disease, Stroke, and Systemic Markers Of C-Reactive Protein and Interleukin-6. *Journal of Periodontology*, 2018, 89(2), 157-165. <https://doi.org/10.1002/JPER.17-0426>
11. Brunese, L., Mercaldo, F., Reginelli, A., Santone, A. An Ensemble Learning Approach for Brain Cancer Detection Exploiting Radiomic Features, *Computer Methods and Programs in Biomedicine*, 2020. <https://doi.org/10.1016/j.cmpb.2019.105134>
12. Cao, Y., Li, P., Zhang, Y. Parallel Processing Algorithm for Railway Signal Fault Diagnosis Data Based on Cloud Computing, *Future Generation Computer Systems*, 2018. <https://doi.org/10.1016/j.future.2018.05.038>

13. Chen, Z., Wu, M., Gao, K., Wu, J., Ding, J., Zeng, Z., Li, X. A Novel Ensemble Deep Learning Approach for Sleep-Wake Detection Using Heart Rate Variability and Acceleration. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2020. <https://doi.org/10.1109/TETCI.2020.2996943>
14. Cios, K.J., Kurgan, L. A. CLIP4: Hybrid Inductive Machine Learning Algorithm That Generates Inequality Rules. *Information Sciences*, 2004. <https://doi.org/10.1016/j.ins.2003.03.015>
15. Cord, A., Chambon, S. Automatic Road Defect Detection by Textural Pattern Recognition Based on AdaBoost. *Computer-Aided Civil and Infrastructure Engineering*, 2012. <https://doi.org/10.1111/j.1467-8667.2011.00736.x>
16. Cumming, G. Replication and P Intervals: P Values Predict the Future Only Vaguely, But Confidence Intervals Do Much Better. *Perspectives on Psychological Science*, 2008, 3(4), 286-300. <https://doi.org/10.1111/j.1745-6924.2008.00079.x>
17. Geurts, P., Ernst, D., Wehenkel, L., Extremely Randomized Trees. *Machine Learning*, 2006, 63(1), 3-42. <https://doi.org/10.1007/s10994-006-6226-1>
18. Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K., Stanley, H. E. PhysioBank, PhysioToolkit, and PhysioNet: Components of A New Research Resource for Complex Physiologic Signals. *Circulation*, 2000, 101(23), e215-e220. <https://doi.org/10.1161/01.CIR.101.23.e215>
19. Guyon, I., Weston, J., Barnhill, S., Vapnik, V. Gene Selection For Cancer Classification Using Support Vector Machines. *Machine Learning*, 2002. <https://doi.org/10.1023/A:1012487302797>
20. Hariharan, R., Thaseen, I. S., Devi, G. U. Performance Analysis of Single-And Ensemble-Based Classifiers for Intrusion Detection. In *Soft Computing for Problem Solving*, Springer, 2019. [https://doi.org/10.1007/978-981-15-0184-5\\_65](https://doi.org/10.1007/978-981-15-0184-5_65)
21. Hu, G., Yin, C., Wan, M., Zhang, Y., Fang, Y. Recognition of Diseased Pinus Trees in UAV Images Using Deep Learning and AdaBoost Classifier. *Biosystems Engineering*, 2020. <https://doi.org/10.1016/j.biosystemseng.2020.03.021>
22. Hung, C., Chen, J. H. A Selective Ensemble Based on Expected Probabilities for Bankruptcy Prediction. *Expert Systems with Applications*, 2009. <https://doi.org/10.1016/j.eswa.2008.06.068>
23. Jabeen, F., Maqsood, M., Ghazanfar, M. A., Aadil, F., Khan, S., Khan, M. F., Mehmood, I. An IoT Based Efficient Hybrid Recommender System for Cardiovascular Disease. *Peer-to-Peer Networking and Applications*, 2019, 12(5), 1263-1276. <https://doi.org/10.1007/s12083-019-00733-3>
24. James, G., Witten, D., Hastie, T., Tibshirani, R. *An Introduction to Statistical Learning*. New York: Springer, 2013. <https://doi.org/10.1007/978-1-4614-7138-7>
25. Jordan, M. I., Mitchell, T. M. Machine learning: Trends, Perspectives, And Prospects, *Science*, 2015, 349(5), 255-260. <https://doi.org/10.1126/science.aaa8415>
26. Kamal, P., Ahuja, S. An Ensemble-Based Model for Prediction of Academic Performance of Students in Undergrad Professional Course. *Journal of Engineering, Design and Technology*, 2019. <https://doi.org/10.1108/JEDT-11-2018-0204>
27. Kamiński, B., Jakubczyk, M., Szufel, P. A Framework for Sensitivity Analysis of Decision Trees. *Central European Journal of Operations Research*, 2018. <https://doi.org/10.1007/s10100-017-0479-6>
28. Khairalla, M.A, Ning, X, Al-Jallad, N.T., El-Faroug, M.O. Short-Term Forecasting for Energy Consumption Through Stacking Heterogeneous Ensemble Learning Model. *Energies*, 2018, 11, 1605. <https://doi.org/10.3390/en11061605>
29. Kotsiantis, S. B., Zaharakis, I., Pintelas, P. Supervised Machine Learning: A Review of Classification Techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, 2007, 160(1), 3-24.
30. Kurgan, L. A., Cios, K. J., Tadeusiewicz, R., Ogiela, M., Goodenday, L. S. Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis. *Artificial Intelligence in Medicine*, 2001, 23(2), 149-169. [https://doi.org/10.1016/S0933-3657\(01\)00082-3](https://doi.org/10.1016/S0933-3657(01)00082-3)
31. Marak, D. C. B., Halder, A., Kumar, A. Semi-supervised Ensemble Learning for Efficient Cancer Sample Classification from miRNA Gene Expression Data. *New Generation Computing*, 2021. <https://doi.org/10.1007/s00354-021-00123-5>
32. Moody, G. B., Mark, R. G. The Impact of the MIT-BIH Arrhythmia Database. *IEEE Engineering in Medicine and Biology Magazine*, 2001, 20(3), 45-50. <https://doi.org/10.1109/51.932724>
33. Muzammal, M., Talat, R., Sodhro, A. H., Pirbhulal, S. A Multi-Sensor Data Fusion Enabled Ensemble Approach for Medical Data from Body Sensor Networks.

- Information Fusion, 2020, 53, 155-164. <https://doi.org/10.1016/j.inffus.2019.06.021>
34. Natekin, A., Knoll, A. Gradient Boosting Machines, A Tutorial. *Frontiers in Neurorobotics*, 2013. <https://doi.org/10.3389/fnbot.2013.00021>
  35. Nilashi, M., Ahmadi, H., Shahmoradi, L., Ibrahim, O., Akbari, E. A Predictive Method for Hepatitis Disease Diagnosis Using Ensembles of Neuro-Fuzzy Technique. *Journal Of Infection and Public Health*, 2019, 12(1), 13-20. <https://doi.org/10.1016/j.jiph.2018.09.009>
  36. Panda D., Dash S. R. Predictive System: Comparison of Classification Techniques for Effective Prediction of Heart Disease. In: *Smart Intelligent Computing and Applications*, 2020. [https://doi.org/10.1007/978-981-13-9282-5\\_19](https://doi.org/10.1007/978-981-13-9282-5_19)
  37. Pinto, A., Pereira, S., Rasteiro, D., Silva, C. A. Hierarchical Brain Tumour Segmentation Using Extremely Randomized Trees. *Pattern Recognition*, 2018. <https://doi.org/10.1016/j.patcog.2018.05.006>
  38. Pławiak, P., Acharya, U. R. Novel Deep Genetic Ensemble of Classifiers for Arrhythmia Detection Using ECG signals. *Neural Computing and Applications*, 2020, 32(15), 11137-11161. <https://doi.org/10.1007/s00521-018-03980-2>
  39. Ponomareva, N., Radpour, S., Hendry, G., Haykal, S., Colthurst, T., Mitrichev, P., Grushetsky, A. TF Boosted Trees: A Scalable TensorFlow Based Framework for Gradient Boosting. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2017. [https://doi.org/10.1007/978-3-319-71273-4\\_44](https://doi.org/10.1007/978-3-319-71273-4_44)
  40. Prakash, V. J., Karthikeyan, N. K. Enhanced Evolutionary Feature Selection and Ensemble Method for Cardiovascular Disease Prediction. *Interdisciplinary Sciences: Computational Life Sciences*, 2021. <https://doi.org/10.1007/s12539-021-00430-x>
  41. Probst, P., Wright, MN, Boulesteix, A.-L. *Hyperparameters and Tuning Strategies for Random Forest*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2019. <https://doi.org/10.1002/widm.1301>
  42. Rath, A., Mishra, D., Panda, G., Satapathy, S. C. Heart Disease Detection Using Deep Learning Methods from Imbalanced ECG Samples. *Biomedical Signal Processing and Control*, 2021. <https://doi.org/10.1016/j.bspc.2021.102820>
  43. Rhanoui, M., Mikram, M., Yousfi, S., Barzali, S. A CNN-BiLSTM Model for Document-Level Sentiment Analysis. *Machine Learning and Knowledge Extraction*, 2019, 1(3), 832-847. <https://doi.org/10.3390/make1030048>
  44. Ricco, Eric Heart Disease Dataset. [Online]. Available: [http://eric.univ-lyon2.fr/%7Ericco/tanagra/fichiers/heart\\_disease\\_male.xls](http://eric.univ-lyon2.fr/%7Ericco/tanagra/fichiers/heart_disease_male.xls) [Accessed in 8 March 2021].
  45. Rokach, L. Ensemble-Based Classifiers. *Artificial Intelligence Review*, 2010, 33(1), 1-39. <https://doi.org/10.1007/s10462-009-9124-7>
  46. Sagi, O., Rokach, L. *Ensemble Learning: A Survey*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018. <https://doi.org/10.1002/widm.1249>
  47. Sahu, S. K., Mohapatra, D. P., Rout, J. K., Sahoo, K. S., Luhach, A. K. An Ensemble-Based Scalable Approach for Intrusion Detection Using Big Data Framework. *Big Data*, 2021, 9(4), 303-321. <https://doi.org/10.1089/big.2020.0201>
  48. Sapra, L., Sandhu, J. K., Goyal, N. Intelligent Method for Detection of Coronary Artery Disease with Ensemble Approach. In *Advances in Communication and Computational Technology*, 2021. [https://doi.org/10.1007/978-981-15-5341-7\\_78](https://doi.org/10.1007/978-981-15-5341-7_78)
  49. Sidiq, U., Aaqib, S. M. An Empirical Model for Thyroid Disease Diagnosis Using Data Mining Techniques. In *International Conference on Sustainable Communication Networks and Application*, Springer, Cham, 2019. [https://doi.org/10.1007/978-3-030-34515-0\\_61](https://doi.org/10.1007/978-3-030-34515-0_61)
  50. Silva, L. O. L. A., Koga, M., Boas, L. B. V., Cugnasca, C. E., Costa, A. H. R. Comparative Assessment of Feature Selection and Classification Techniques for Visual Inspection of Pot Plant Seedlings. *Computers And Electronics in Agriculture*, 2013. <https://doi.org/10.1016/j.compag.2013.07.001>
  51. Siwek, K., Osowski, S. Improving the Accuracy of Prediction of PM10 Pollution by The Wavelet Transformation and An Ensemble of Neural Predictors. *Engineering Applications of Artificial Intelligence*, 2012, 25(6), 1246-1258. <https://doi.org/10.1016/j.engappai.2011.10.013>
  52. Surakhi, O. M., Zaidan, M. A., Serhan, S., Salah, I., Hussein, T. An Optimal Stacked Ensemble Deep Learning Model for Predicting Time-Series Data Using a Genetic Algorithm-An Application for Aerosol Particle Number Concentrations. *Computers*, 9(4), 89, 2020. <https://doi.org/10.3390/computers9040089>
  53. Tama, B. A., Im, S., Lee, S. Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble. *BioMed Research International*, 2020. <https://doi.org/10.1155/2020/9816142>

54. Tan, K. K., Le, N. Q. K., Yeh, H. Y., Chua, M. C. H. Ensemble of Deep Recurrent Neural Networks for Identifying Enhancers Via Dinucleotide Physicochemical Properties. *Cells*, 2019, 8(7), 767. <https://doi.org/10.3390/cells8070767>
55. Uslu, S. Optimization of Diesel Engine Operating Parameters Fueled with Palm Oil-Diesel Blend: Comparative Evaluation Between Response Surface Methodology (RSM) And Artificial Neural Network (ANN). *Fuel*, 2020. <https://doi.org/10.1016/j.fuel.2020.117990>
56. Wang, S. J., Mathew, A., Chen, Y., Xi, L. F., Ma, L., Lee, J. Empirical Analysis of Support Vector Machine Ensemble Classifiers. *Expert Systems with Applications*, 2009. <https://doi.org/10.1016/j.eswa.2008.07.041>
57. WHO, Cardiovascular Diseases in India. [Online]. Available: <https://www.who.int/india/health-topics/cardiovascular-diseases> [Accessed in 8 March 2021].
58. WHO, Cardiovascular Diseases. [Online]. Available: <https://www.who.int/health-topics/cardiovascular-diseases> [Accessed in 8 March 2021].
59. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., Steinberg, D. Top 10 Algorithms in Data Mining. *Knowledge Information Systems*, 2010.
60. Wu, Z., Shi, L., Li, J., Wang, Q., Sun, L., Wei, Z., Plaza, J., Plaza, A. GPU Parallel Implementation of Spatially Adaptive Hyperspectral Image Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2017, 11(4), 1131-1143. <https://doi.org/10.1109/JSTARS.2017.2755639>
61. Xie, Q., Cheng, G., Xu, X., Zhao, Z. Research Based on Stock Predicting Model of Neural Networks Ensemble Learning. In *MATEC Web of Conferences, EDP Sciences*, 2018. <https://doi.org/10.1051/mateconf/201823202029>
62. Xu, S. Bayesian Naïve Bayes Classifiers to Text Classification. *Journal of Information Science*, 2018. <https://doi.org/10.1177/0165551516677946>
63. Yekkala, I., Dixit, S. A Novel Approach for Heart Disease Prediction Using Genetic Algorithm and Ensemble Classification. *Proceedings of SAI Intelligent Systems Conference*, 2021, 468-489. [https://doi.org/10.1007/978-3-030-55187-2\\_36](https://doi.org/10.1007/978-3-030-55187-2_36)
64. Zeng, N., Qiu, H., Wang, Z., Liu, W., Zhang, H., Li, Y. A New Switching-Delayed-PSO-Based Optimized SVM Algorithm for Diagnosis of Alzheimer's Disease. *Neurocomputing*, 2018. <https://doi.org/10.1016/j.neucom.2018.09.001>
65. Zhang, X., Waller, S. T., Jiang, P. An Ensemble Machine Learning-Based Modeling Framework for Analysis of Traffic Crash Frequency. *Computer-Aided Civil and Infrastructure Engineering*, 2019. <https://doi.org/10.1111/mice.12485>
66. Zhao, J., Feng, Q., Wu, P., Lupu, R. A., Wilke, R. A., Wells, Q. S., Denny, J. C., Wei, W. Q. Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction. *Scientific Reports*, 2019, 9(1), 1-10. <https://doi.org/10.1038/s41598-018-36745-x>
67. Zhenya, Q., Zhang, Z. A Hybrid Cost-Sensitive Ensemble for Heart Disease Prediction. *BMC Medical Informatics and Decision Making*, 2021, 21(1), 1-18. <https://doi.org/10.1186/s12911-021-01436-7>
68. Zhou, X., Li, Y., Liang, W. CNN-RNN Based Intelligent Recommendation for Online Medical Pre-Diagnosis Support. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020. <https://doi.org/10.1109/TCBB.2020.2994780>

