Applied Computer Science, vol. 19, no. 1, pp. 107–132
doi: 10.35784/acs-2023-08

Submitted: 2023-02-21 | Revised: 2023-03-07 | Accepted: 2023-03-14

Keywords: Classification, Naïve Bayes, Neural Network, Support Vector Machine, Decision Tree, Ensemble Learning, Random Forest

Archana GUNAKALA [0000-0002-3375-1893]*, *Afzal Hussain SHAHID*[**]

A COMPARATIVE STUDY ON PERFORMANCE OF BASIC AND ENSEMBLE CLASSIFIERS WITH VARIOUS DATASETS

Abstract

Classification plays a critical role in machine learning (ML) systems for processing images, text and high -dimensional data. Predicting class labels from training data is the primary goal of classification. An optimal model for a particular classification problem is chosen based on the model's performance and execution time. This paper compares and analyzes the performance of basic as well as ensemble classifiers utilizing 10-fold cross validation and also discusses their essential concepts, advantages, and disadvantages. In this study five basic classifiers namely Naïve Bayes (NB), Multi-layer Perceptron (MLP), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) and the ensemble of all the five classifiers along with few more combinations are compared with five University of California Irvine (UCI) ML Repository datasets and a Diabetes Health Indicators dataset from Kaggle repository. To analyze and compare the performance of classifiers, evaluation metrics like Accuracy, Recall, Precision, Area Under Curve (AUC) and F-Score are used. Experimental results showed that SVM performs best on two out of the six datasets (Diabetes Health Indicators and waveform), RF performs best for Arrhythmia, Sonar, Tic-tac-toe datasets, and the best ensemble combination is found to be DT+SVM+RF on Ionosphere dataset having respective accuracies 72.58%, 90.38%, 81.63%, 73.59%, 94.78% and 94.01%. The proposed ensemble combinations outperformed the conventional models for few datasets.

1. INTRODUCTION

Machine Learning is having a significant impact on a wide variety of applications, including text comprehension, image, speech recognition, health care, anomaly detection and more. An ML algorithm receives a set of data known as training data as input. It looks for patterns in the input data and trains the model to produce predicted outputs (target) (Pugliese et al., 2021). Once the ML model is trained, it provides the output for the new test data. The goal or label is the value that the ML model must predict. A labeled dataset contains number of samples and features of data, along with a class label that is required for model building. ML provides a variety of classification techniques for predicting class labels.

[*] VIT-AP University, Amaravati-522237, Andhra Pradesh, India, archu.gunakala@gmail.com
[**] VIT-AP University, Amaravati-522237, Andhra Pradesh, India, syedahshahid@gmail.com

The task of classification is to learn a function (F) also called the target function, that associates each of the sets of attributes $X = \{X_1, X_2, X_3,,,., X_n\}$ with an associated class (Y). The ML model and its parameters, as well as the features extracted and chosen as inputs, all play a major role in the classification performance (Alshayeji et al., 2022). It can be difficult to choose the appropriate classification algorithm among the basic as well as ensemble classifiers available. The classifier is determined on the basis of data sets and desired outcomes. In this paper, the performance of basic supervised machine learning classifiers such as NB, SVM, MLP, DT, and RF as well as heterogeneous ensemble classifiers is evaluated on five UCI datasets and a Diabetes Health Indicator dataset from Kaggle. Each of the basic as well as ensemble classifier's performances with each dataset is calculated using metrics such as accuracy, recall, precision, f-score, and AUC.

The remaining part of this paper is organized as follows: A summary of the relevant literature is given in Section 2. Section 3 describes the different classifiers employed and discusses their advantages and disadvantages. Section 4 discusses ensemble methods and their classification techniques. Section 5 describes the methodology. Section 6 represents the results and discussion. And finally, the conclusion of the paper is given in Section 7.

2. LITERATURE REVIEW

Machine Learning is a broad area in AI that studies how to construct learning-capable systems (Ganie & Malik, 2022). Research in ML focuses on the construction of accurate and efficient classifiers for large databases. In (Ma et al., 2020) authors have compared NB and SVM to classify whether an email is spam or a ham based on the content of the email, and the results showed that as the size of the training data increased, the accuracy of both classifiers also increased, but SVM showed higher accuracy. To prevent the cognitive inabilities and predict early prediction of Alzheimer's in elderly people, the paper (Revathi et al., 2022) proposed a two-stage classification with SVM (86%) and RF (71%), first with the diabetes and hypertension symptoms**,** and then applied multinomial logistic regression (89%) to the results of cognitive ability test and classified the risk as "No Alzheimer", "Uncertain Alzheimer's" and "Definite Alzheimer's". A comparison of different models in classification, data mining, ML, and deep learning for the prediction of Cardio-vascular diseases (CVD) in (Swathy & Saruladha, 2021) using 3-fold cross-validation and different datasets and tools used for CVD prediction are explored.

For the classification of breast cancer utilizing the Wisconsin breast cancer dataset (WBCD) and the Wisconsin diagnostic breast cancer dataset (WDBC), an ANN model with one hidden layer was employed in (Alshayeji et al., 2022) and had an average accuracy of 99.85% for WBCD and 99.47 % for WDBC. In (Kilincer et al., 2021), the authors performed a detailed classification on different Intrusion Detection datasets, using different classification algorithms like SVM, KNN, DT algorithms and evaluated the performance of max-min normalization using cross validation with 10 folds. The results obtained showed that DT is more successful ranging from 99% to 100% than the remaining classifiers used for all the datasets.

Another approach in (Mohamed, 2017) compared four ML techniques Decision Tree, KNN, ANN and SVM on German Credit Data and resulted in 70-75% of performance and concluded that there is no particular classifier that can satisfy all the criteria. Using the KNN classification technique, a model capable of automatically recognizing iris species is developed in (Thirunavukkarasu et al., 2018) with 100 % accuracy for two classes and a 4%

error rate for one class in the iris dataset. DT and ANN algorithms outperform other algorithms by 91.45% and 91.17%, respectively, in a comparative study conducted by the on an ionosphere dataset using five different classification algorithms, including Naive Bayes, SVM, ANN, K-NN, and J48.

However, most of the current research on classification is being done by using ensemble methods to improve the performance of base classifiers. These methods can be classified into homogeneous (combining similar classifiers-Bagging and Adaboost (AB) and heterogeneous methods (combining different classifiers-Stacking). The authors of a recent study (Alshdaifat et al., 2021) reduced the basic classifiers with low performance such that only the best classifiers remain in the ensemble by considering AUC to determine the efficacy of a classifier. An intelligent ensemble of auto ML system using greedy approach was proposed in (Consuegra-ayala et al., 2022) to select the base models and produced more generalized results than the basic models. New ensemble-based framework, RF, Bagged Decision Tree (BDT), and Extra Tree (ET) with Bagging method and AB, Stochastic Gradient Boosting (SGB) with Boosting method and finally LR, SVM, DT with voting method has been used to predict diabetes along with k-fold cross validation (Ganie & Malik, 2022) and achieved an accuracy of 99.41% with bagged DT. An ensemble learning method using a stacking algorithm with MLP, RF as base classifiers and LR as meta classifier is proposed in (Yakut & Bolat, 2022) along with feature extraction and achieved 99.8% and 99.2% for category and patient based arrhythmic heartbeat classification datasets, respectively.

From the literature review it is evident that most of the researchers have applied either the basic classification algorithms or homogeneous ensemble classifiers. However, this paper investigates the performance of heterogeneous ensemble classifiers on different datasets to show the robustness of the ensemble classifiers. The ensemble classifiers can enhance the strengths of basic classifiers by potentially overcoming the weaknesses of basic classifiers as each basic classifier has its own biases and strengths.

3. CLASSIFICATION

Classification refers to any situation in which a class label must be predicted from given data. The objective is to generalize known structures for application to new data. Fig. 1 illustrates the basic concept of a classification algorithm. Based on the classes present in the data used, classification can be categorized as follows:

Binary Classification

This classification divides input data into one of two groups. Commonly, one of the classes represents a ‘normal/desired’ state, while the other represents an ‘abnormal/undesired’ state.

Multi-Class Classification

The data set is classified into one of several possible classes in multi-class classification. A multi-class algorithm, as opposed to a binary classification, is trained with data that can be classified into one out of several possible classes.

Multi-label Classification

In contrast to binary and multi-class classification, where the outcome belongs to only one class, the multi-label output relates to one or more classes. As a result, the same input data may be classified into multiple classes.

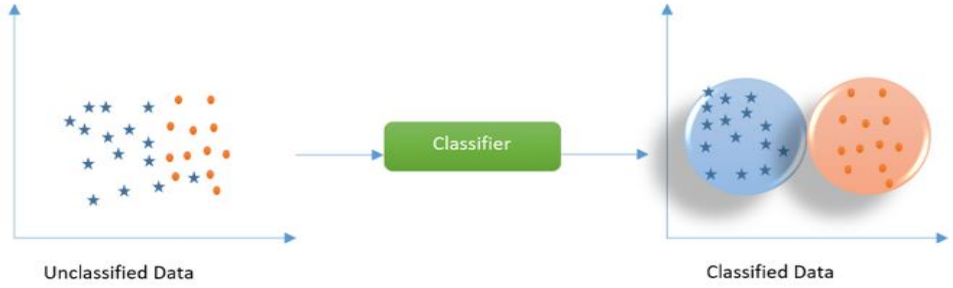

Fig. 1. Basic concept of a classification

3.1. Supervised Learning

The vast majority of actual applications of machine learning make use of supervised learning. In supervised learning, an algorithm learns the mapping function Y = f(X) from input to output Y with input and output variables (X, Y). In order to produce predictions, an algorithm is trained on a known dataset (the training dataset) using a known set of inputs (referred to as features) and known responses (targets). Input data is labeled to match desired outputs or response values, and this forms the basis of the training dataset. For each new dataset, supervised learning builds an algorithmic model by detecting correlations between features and target. The goal of this study is to uncover major trends in the performance and application of several supervised machine learning algorithms on various datasets. Based on the characteristics of our dataset, we selected a subset of the available ML classification approaches.

3.1.1. Naïve Bayes

The NB classifier works on the basis of Bayes rule, in which the conditional independence of the constituent features of X= X_1, X_2,..., X_n, with regard to one another given the outcome Y ={Y_1, Y_2,…., Y_k} (Wade et al., 2017). The Bayes rule can be given as:

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)} \quad (1)$$

Where, Y is the occurrence for which we wish to determine the probability, and X is the new data that is associated to Y. P(Y|X) = posterior probability, which we are attempting to figure out. P(X|Y) = likelihood, which is the probability of discovering new evidence with the given initial hypothesis. P(Y) = prior, which is the probability of our hypothesis in the absence of further prior information. P(X) = marginal likelihood, is the total probability of observing the evidence.

Bayes Theorem and the premise of independent predictors are the foundations of this classifier. There are many variants of NB classifier. First is the Bernouli's NB which is used for binary data, second is Multinomial NB to classify text data in applications like Natural Language Processing (NLP) and the third is Gaussian NB for data with continuous values. In this paper Gaussian NB is used which can be formulated as shown in equation (2) where σ^2 and μ are standard deviation and mean respectively. The NB classifier is easy to develop and useful for extremely big data sets.

$$P(X = x_i|Y = y_k) = \left(2\pi\sigma_{y_k}^2\right)^{-1/2} \exp\left(\frac{\left(\mu_{y_k} - x_i\right)^2}{2\sigma_{y_k}^2}\right) \quad (2)$$

The advantages of NB classifier are as follows: a huge dataset can benefit from the simplicity and efficiency of predicting the class of a test set. When the assumption of independence is true, a Naive Bayes classifier improves the performance over other models and requires less training data. It shows better performance with categorical input data than numerical input data. Numerical variables are assumed to have a normal distribution (bell curve, which is a strong assumption). In addition, it can also be used to address binary and multi-class classification issues (Uddin et al., 2019). It can work with both discrete and continuous data and can make probabilistic predictions. However, NB classifier has some disadvantages. These are: the model will assign 0 (zero) probability and be unable to predict if a categorical variable in the test data set has a category that isn't present in the training data set. Naive Bayes assumes that the features are independent. In real life, this is rare. Although NB learns quickly and efficiently, it might be highly biased if the training set is not optimal.

3.1.2. Support vector Machine

In the year 1963, Alexey Y. Chervonenkis and Vladimir N. Vapnik (Farhat, 1992) developed SVM which can classify linear data as well as non-linear data. Each sample of linear data with n features is first viewed as an n-dimensional feature space. Finding the hyperplane that separates the data into two groups will enhance the marginal distance between the classes while reducing the probability of classification error. The hyperplane with the maximum marginal distance is represented in Fig. 2(a) in 2-dimensional feature space and Fig. 2(b) represents the same in 3-dimensional feature space. If non-linear data is to be classified then SVM uses a special feature called the kernel function, a problem that can be expressed as the quadratic programming problem (Qian et al., 2021), in which a lower-dimensional input space is transformed into a higher-dimensional input space which can be given as:

$$\min_{w,b,\varepsilon} \frac{\| w \|^2}{2} + C \sum_{i=1}^{n} \varepsilon_i \quad (3)$$

where w is the hyperplane's normal to optimal decision, bias term b, and the distance of a hyperplane from its origin can be expressed as $\frac{b}{\|w\|}$, balance between margin and training error is achieved by setting the regularization parameter $C > 0$, and training cases can be misclassified when using $\mathcal{E}_i$'s as weak variables. The kernel function chosen has a direct impact on the optimal solution of the problem. There are several extensions of SVM which allow to solve multiclass classification problems as well as detection of important features (Baumann et al., 2019). The Gaussian kernel function is usually considered. In this study a standard SVM is considered with linear function to compare with other classification models. The robustness of SVM lies in its advantages, which can be given as follows: it functions extremely well with a good margin of separation. It's a good fit for environments with high dimensions. SVM is more advantageous when the attributes are more than the samples. Suitable for balanced datasets. However, it has some disadvantages such as, it

performs poorly when dealing with large data sets, since it requires more time to train. When the data contains more noise, such as target classes that overlap, it performs poorly (Rezvani & Wang, 2022). Cannot perform well for imbalanced datasets.

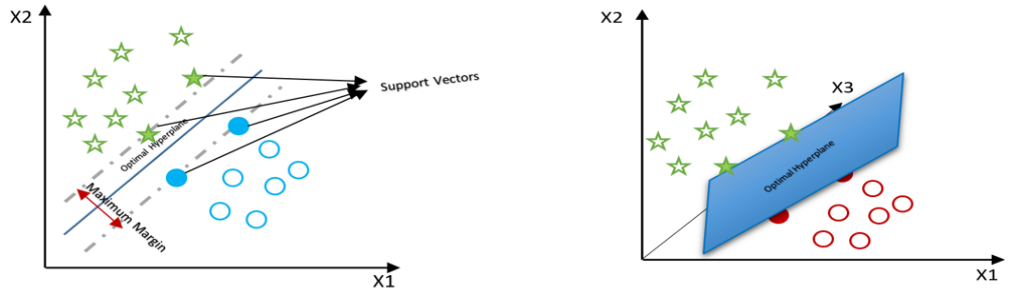

Fig. 2. Optimal hyperplane: a) in 2D feature space; b) in 3D feature space

3.1.3. Neural Networks

A Neural network, also known as artificial neural network is a collection of artificial neurons interconnected with each other in which each neuron works as a mathematical operator which resembles the functionality of a biological neuron. An input layer, a number of hidden layers, and an output layer constitute a neural network. Each hidden layer contains many number of neurons and each neuron in a hidden layer is connected to other layers (either next hidden layer or to the output layer) through links called edges. Each edge is assigned with a weight. Each neuron has it's inputs (the weights associated with the incoming edges and the bias) and output by some activation function (Wei et al., 2022) which can be given as shown in equation (4), where y_k is the output of neuron k, w_{ki} is the weight of the edge connecting the input x_i , the bias term b, and f is the activation function which is also represented in Fig. 3. Finally, the neurons in the output layer indicate the final outcome. MLP is the mostly used neural network model which has an architecture of fully connected neural network. MLP models are trained with back propagation algorithm which is based on error-correction technique mean squared error (MSE) (Fath et al., 2020). If the MSE is larger, the backpropagation operation is continued until the MSE is minimized.

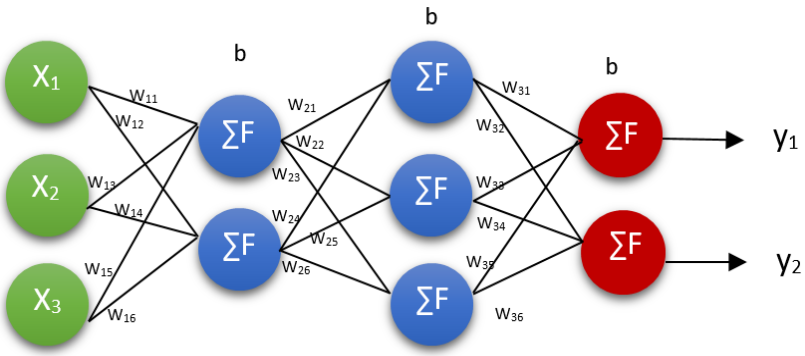

Fig. 3. Artificial Neural Network

Where:

$$y_k = f(\sum_{i=1}^{n} w_{ki} x_i + b) \quad (4)$$

Nonlinear correlations between independent and dependent variables can be detected with this method. Less formal training in statistics is necessary. The ability to choose from a wide range of training methods. Regression and classification problems both can be solved with Neural Networks. However, the disadvantages of Neural Networks are as follows: possess "black box" properties. The user is not exposed to the precise decision-making process. Training a network for a difficult classification problem is computationally intensive and Preprocessing is required for predictor variables and independent variables.

3.1.4. Decision Tree

Decision Tree classifier is a type of Supervised Machine Learning method in which a tree structure is constructed based on the input features. The two entities that can be used to explain the tree are decision nodes from which the data is divided and leaves which represents the decisions or conclusions (Maniruzzaman et al., 2019). A prediction for the outcome of a new observation is made by first determining which leaf it corresponds to, and then integrating the results of the current observations inside that leaf to get a predicted result. The execution is halted if the data split does not yield any benefits (Patel & Prajapati, 2018).

Tab. 1. Decision tree algorithms and their mechanisms

Algorithm	Classification Basis or metrics Used	Mechanism
Classification and regression trees (CART)	Gini Index	Utilizes numerical splitting to form a tree (Jia & Qiu, 2020).
Iterative Dichotomiser 3 (ID3)	Entropy and Information gain	Continuous datasets include only discrete values, so they are labeled as discrete datasets (Priyanka & Kumar, 2020).
C4.5	Improvised version of ID 3	Adapts to both a discrete and a continuous dataset. In addition, it is able to handle datasets that are not complete. "PRUNNING" is a strategy that addresses the issue of over filtering.
C5.0	Improvised algorithm of C4.5	C5.0 algorithm allows to select between predicting missing values based on other attributes or statistically distributing the case between outcomes.
Chi-square Automatic Interaction Detector (CHAID)	This version predates the ID3 implementation	Dependent variables are detected from the categorized variables of a dataset and is used for nominal scaled variable(Punyapornwithaya et al., 2022).

Table 1 depicts the decision tree algorithms used to split the attributes to test at each node and to assess whether the attribute splitting is optimal for individual classes. Because the splitting criteria must be same, the resultant partitions at each branch is as PURE (belonging to the same class) as possible. The decision tree algorithms mentioned in Table 1 make use of some metrics to consider an attribute as a decision node out of the available attributes in a dataset while constructing a decision tree. These metrics are also called "attribute selection measures" and are used to reduce impurities. The higher the impurity reduction, the better the split attribute chosen (Patel & Upadhyay, 2012). The attribute selection measures can be given as follows:

a) Entropy:

A metric for estimating the amount of impurity in a group of data is called entropy. It indicates how the data is split and the quality of spilt in a decision tree. For given *S* samples and *N* features with probability $p(x_i)$, i = 0 to N, the entropy can be calculated as in equation (5).

$$H(S) = \sum_{i=0}^{N} log_2 \left(\frac{1}{p(x_i)}\right)^{p(x_i)} \quad (5)$$

b) Information Gain:

Information gain, also known as Kullback-Leibler divergence, is the entropy of a dataset *S* after it has been segmented depending on an attribute *A* is shown in equation (6).

$$Info\ Gain(S, A) = H(S) - \sum_{i=0}^{N} p(x_i) \times H(x_i) \quad (6)$$

Information Gain indicates the amount of data a feature provides about a class. In order to build the decision tree, we divide each node in half based on the information it gains. A decision tree technique splits the node or attribute with the largest information gain first.

c) Gain Ratio:

The information gain measure is biased in favor of tests with a lot of results. As a result, attribute-based partitioning produces the most information, but is poor for categorization (Patel & Upadhyay, 2012). ID3's successor, C4.5, utilizes a Gain ratio extension to the information gain. The gain ratio can be formulated as in equation (7):

$$Gain\ Ratio = \frac{Information\ Gain}{Entropy} \quad (7)$$

d) Gini Index:

It is a measure used to determine the impureness of a dataset feature i.e., how well a DT was split. Calculations were made to determine the probability that a randomly chosen feature would be incorrectly classified. The Gini Index has a range of 0 to 1, with 0 denoting classification purity and 1 denoting random distribution of elements over different classes. In (Kushwah et al., 2021) this is applied to construct a decision tree using CART algorithm. The CART method generates a decision tree with the use

of a binary split by using the Gini Index. The Gini Index can be represented as shown in equation (8).

$$GI = \sum_{i=0}^{n}(p_i - {p_i}^2) \quad (8)$$

The advantages of decision tree classifier are as follows: The final classification tree is easy to interpret and understand, Data preparation is simple, All the three kinds of data – numeric, categorical and nominal – can be classified and can provide strong classifiers which can be tested through the use of statistics. However it has some disadvantages, they are: DT assumes that each class is mutually exclusive with the others, the algorithm cannot split if the attribute value of a non-leaf node is absent and the algorithm is determined by the sequence of variables or attributes.

4. ENSEMBLE LEARNING

Ensemble techniques analyze a broad variety of models and aggregate them to generate a single final model rather than generating just one model and hoping that it is the finest accurate prediction we can create. Ensemble learning improves machine learning outcomes by incorporating numerous competing models. In comparison to a single model, this approach is more accurate at predicting future outcomes (Nazari et al., 2021). There are two types of ensemble methods: homogeneous ensemble methods and heterogeneous ensemble methods (Tewari & Dwivedi, 2020).

Homogeneous Ensemble methods

In the same way that the Random Forest model was built, a homogeneous ensemble is a collection of classifiers that were produced using a different subset of data. Homogeneous ensembles include bagging, random forest and random subspace.

Heterogeneous ensemble methods

Heterogeneous ensemble is a collection of various classifiers developed from the same data. Voting, Stacking are the examples.

4.1. Random Forest

Collections of decision trees are known as random forests. Combining the results of multiple predictors is a common type of ensemble approach. In addition to this, the bagging technique employed by random forest is used, which allows each tree to be trained on a random sample of an original dataset and gets an overall consensus. On the other hand, decision trees are more interpretable since they have fewer levels of complexity. Several other classifiers, such as AB, SVM, NN, and DT are less accurate because they overfit. Another application of this technique is as a means of selecting features based on their perceived importance (Shafi et al., 2020).

Both regression and classification problems can be solved with random forests. For a given dataset D(X,Y), the training data $X=\{x_1,x_2,\ldots..x_n\}$ with their respective class labels $Y=\{y_1,y_2,\ldots..y_n\}$ where n is the total number of samples in the training dataset, like Bagging technique random forest considers replacing the training set, with a random sample, repeatedly for B times ,say,and fits trees to these samples. i.e, If X_b, Y_b are n training instances sampled with replacement from X and Y, Random forest trains a classification or

regression tree (f_b) on these instances where b=1, 2... B. After training the model, in case of regression, the predictions for testing samples $\hat{x}$ are given by averaging the predictions of all the individual regression trees as shown in equation (9), whereas in case of classification, testing sample predictions are considered by the majority voting technique as shown in Fig. 4.

$$\hat{f} = \frac{\sum_{b=1}^{B} f_b(\hat{x})}{B} \tag{9}$$

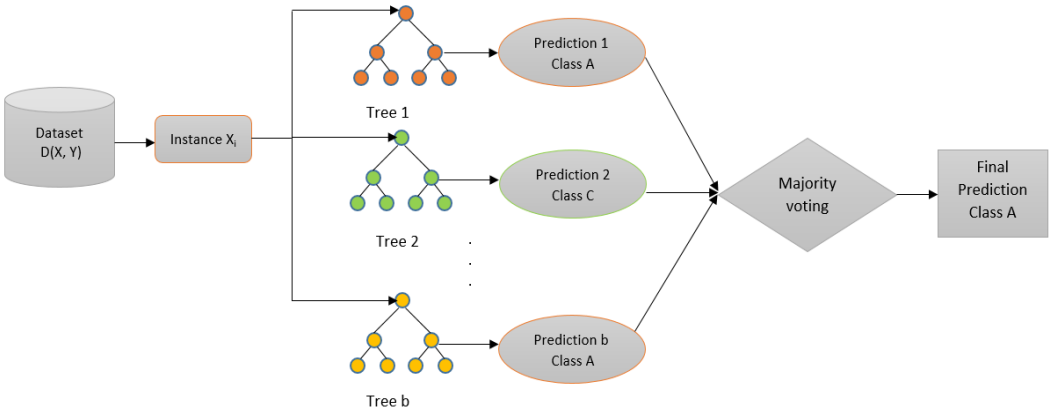

Fig. 4. Prediction Strategy of random Forest for classification

The advantages of random forest algorithm are as follows: It improves accuracy by reducing decision tree overfitting, It can be used for classification and regression difficulties, We can use either continuous or categorical values in random forest, Filling up missing data values is done automatically, Since a rule-based approach is employed, normalization is not required, Random Forest is less affected by noise. However, there are some disadvantages of RF. They are: It consumes a lot of resources and computational power because it produces many trees to aggregate their outputs and Training is lengthy since it uses a variety of decision trees to select a class.

5. METHODOLOGY

In this paper, each considered dataset is classified with different basic classifiers as well as few ensembles of basic classifiers. In order to evaluate the models a 10-fold cross validation is performed on each dataset and the predictions are given. The workflow of the ensemble combination of classifiers is given in Fig. 5 where, P1, P2, P3, P4, P5 are the predictions obtained from DT, MLP, NB, SVM, RF respectively.

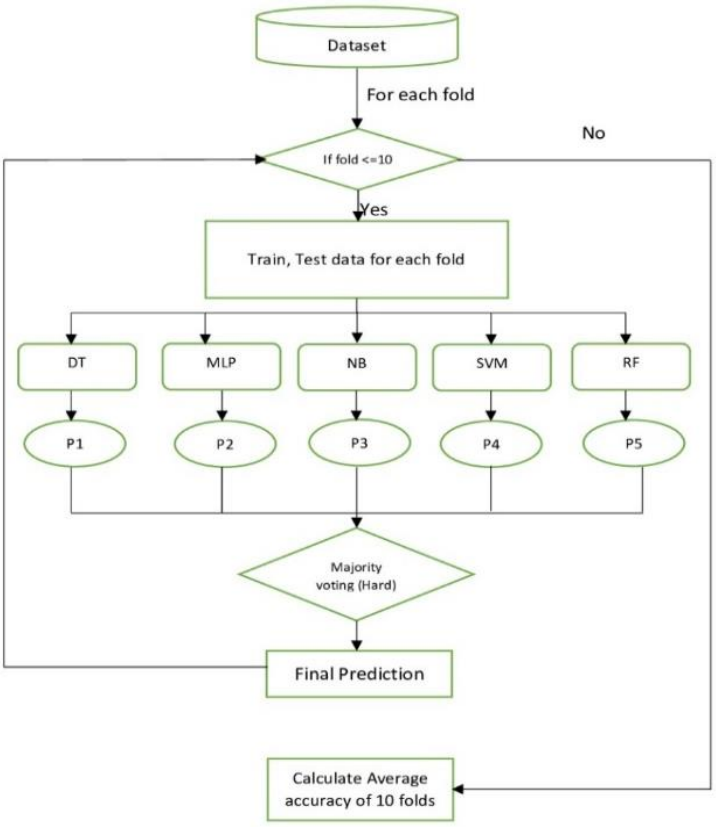

Fig. 5. Work flow of ensemble of all the basic classifiers

All the predictions obtained from these basic classifiers are combined based on the majority voting (hard) technique in each fold of the 10-fold cross validation and a final prediction will be generated. The average of accuracies obtained for each fold are calculated and a final accuracy of that model is generated. The evaluation metrics used in order to evaluate the models and the datasets considered are described in the following subsections.

5.1. Scheme of methodology

An overview of the approach is shown schematically in Fig. 6. At first, the literature is studied and the existing methods used by most of the researchers are analyzed. Then, based on the literature, there are research gaps, such as the fact that most research is using only homogeneous ensemble classifiers or some basic machine learning classifiers. Over and above, existing works applied their classification models only to a few datasets. Therefore, this study includes five basic classifiers, each of which is applied individually on different datasets, and proposed a new model by combining the classifiers using heterogeneous ensemble techniques and applying them to the datasets with hard majority voting. Finally, the results of both basic classifiers as well as heterogeneous ensemble classifiers are compared with the existing works.

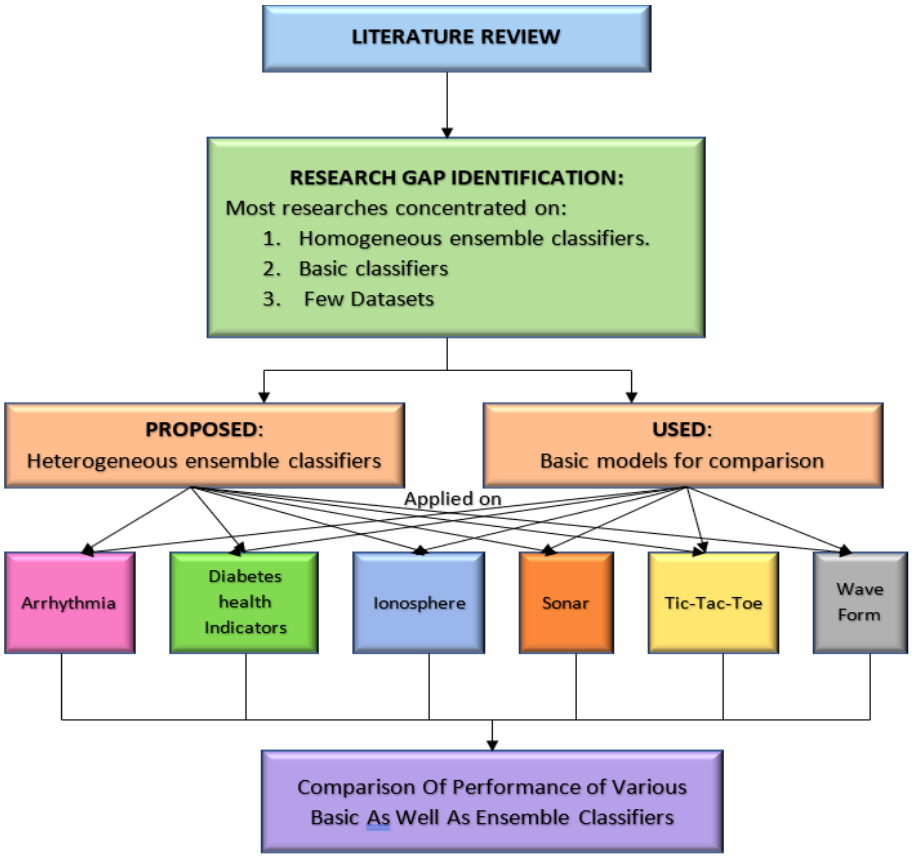

Fig. 6. Scheme of Research Methodology

5.2. Dataset Description

The datasets considered for the comparative study were taken from the UCI repository and Kaggle due to their popularity in existing works and the prevalence of certain attributes in the database (combination of attribute types, binary and multi-class labels, and small and large sizes). The selection of datasets is also based on the wide range of applications like signal data, game theory, and medical data. Also, as they are popularly used by some of the existing works Arrhythmia, Diabetes Health Indicators, Ionosphere, Sonar, Tic-Tac-Toe and Wave Form datasets are considered to compare the performance of the proposed model with the existing works. The characteristics of each dataset are mentioned in Table 2 with varying sizes and dimensions.

The dataset Arrhythmia is used to classify cardiac arrhythmia in one of the 16 groups. Class 1 indicates 'normal' i.e., absence of cardiac arrhythmia, class 2 to class 15 indicates various ECG classes of arrhythmia and class 16 indicates the remaining groups of Arrhythmia. The dataset consists of 279 attributes of which 206 have linear values and the remaining are nominal values.

The Behavioral Risk Factor Surveillance System (BRFSS) 2015 survey received 70,692 replies, and these responses make up the Diabetes Health Indicators dataset from the Kaggle repository. Samples without diabetes and those with either prediabetes or diabetes are split

in equal ratios in the dataset. The target variable is divided into two groups with 0 representing no diabetes and 1 representing prediabetes or diabetes and this diabetic health indicators dataset is a balanced dataset comprising of 21 feature variables.

Ionosphere dataset is the classification of radar signals from ionosphere. The received radar signals are analyzed using an autocorrelation function with the parameters pulse time and pulse number. Consists of 351 samples described by two attributes per pulse number. It has 34 attributes of continuous type and two target classes in which 0 if for bad radar return and 1 for good radar return indicating some type of structure in Ionosphere.

The dataset Sonar Consists of 208 samples of mine (metal cylinder) and rocks obtained by bouncing sonar signals on a metal cylinder at different angles spanning upto90 degrees (for mine) and 180 degrees (for rocks). Each sample is a pattern of 60 numbers (attributes) ranging from 0.0 to 1.0 and it is a binary class classification dataset in which each record is labeled 1, 2 for Rock and Mine classes.

Tab. 2. Characteristic summary of datasets

Dataset	Samples	Attributes	No. of classes
Arrhythmia	452	279	16
Diabetes Health Indicators	70692	21	2
Ionosphere	351	34	2
Sonar	208	60	2
Tic-tac-toe	958	9	2
Wave Form	5000	40	3

Tic-tac-toe is a binary classification dataset with 958 samples representing all the possible combinations of tic-tac-toe game. Consists of 9 feature variables representing the position of each cell in a 3×3 box in the game and the target labels are 0 (for losing the game) and 1 (for winning the game).

The Waveform is a balanced multiclass classification dataset consisting of 5000 samples, 40 attributes and 3 classes each of which is generated from a combination 2 of 3 base waves.

5.3. Evaluation Metrics

It is not our goal to create a predictive model. Out-of-the-sample data is the key to building and selecting the best model. The validation of the model is therefore essential before computing predictive values. A predictive model's performance is quantified by the evaluation metrics, in model building, picking the proper statistical metric is critical because the measures used have an impact on how machine learning algorithm performance is evaluated (Fang et al., 2022) and compared and also have an impact on consideration of various characteristics in the results, as well as final decision on algorithm to be used. There were numerous statistical indicators to investigate for classification problems.

5.3.1. Confusion Matrix

Tab. 3. Confusion Matrix

ACTUAL CLASS	PREDICTED CLASS		
		P_{ve}	N_{ve}
	P_{ve}	TP	FN
	N_{ve}	FP	TN

A matrix with N × N where N is no. of predicted classes is called confusion matrix. Classification algorithms are frequently evaluated using this method (Sevinç, 2022). When there is a significant difference between the classes, it is used. Table 3 displays a confusion matrix for N = 2, with the following interpretations for the entries where P_{ve} indicates positive and N_{ve} indicates negative.

True Positive (TP): Those values in which the classified values and the actual values both are positive.

True Negative (TN): Those values in which the classified values and the actual values both are negative.

False Positive (FP): Values that were actually negative but were predicted to be positive, which is called Type I Error.

False Negative (FN): Values that were classified as negative but were actually positive. Which is also called Type II Error.

5.3.2. Accuracy

The accuracy statistic represents total predictive performance, indicating how many correct predictions are made. Accuracy can be represented as:

$$\textbf{Accuracy} = \frac{\textbf{TP} + \textbf{TN}}{\textbf{TP} + \textbf{TN} + \textbf{FP} + \textbf{FN}}$$

If the datasets are symmetric (the count of false negative values and false positive values are nearly equal) and the costs of false negative values and false positive values are equal, accuracy metric can be a helpful indicator. In this paper, cross validation of 10-folds is used and hence the average accuracy of all the 10 folds is considered as the evaluation metric.

5.3.3. Precision

A measure of precision reveals how many of the positive predictions were right out of total predictions that are positive. For an unbalanced dataset, precision calculates the accuracy of the minor class.

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FP}}$$

5.3.4. Recall

The recall metric which is also called sensitivity shows the number of correct positive predictions out of actual positive cases were made.

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FN}}$$

5.3.5. F-Score

The accuracy of a model can be measured using the metric F-score, commonly known as the F1-score. Binary classification systems classify examples as "positive" or "negative" using this metric. The F-Score can also be measured as the harmonic mean of the precision & recall of a model. The F-score can be changed such that recall takes precedence over precision, and vice versa. Common modified F-scores include the F0.5-score, the F2-score, and the standard F1-score. The F-score of a perfect model is 1 in which the contribution of precision and recall are same. It can be stated numerically as follows:

$$\textbf{F} - \textbf{Score} = \frac{\textbf{2TP}}{\textbf{2TP} + \textbf{FP} + \textbf{FN}} = \textbf{2} \times \frac{\textbf{Precision} \times \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}}$$

5.3.6. Area Under Curve

The receiver operating characteristic (ROC) curve's AUC metric displays the overall accuracy of both positive and negative predictions, but it will handle imbalanced problems better than the accuracy metric.

$$\textbf{AUC} = \frac{\left(\frac{\textbf{TP}}{\textbf{TP} + \textbf{FN}} + \frac{\textbf{TN}}{\textbf{TN} + \textbf{FP}}\right)}{\textbf{2}}$$

6. RESULTS AND DISCUSSION

This experiment was carried out using python 3.10.2 version in the Windows (10) environment. Configuration of computer hardware environment is the following: the system is 64-bit with Windows10 operating system, the processor is Intel (R) core (TM) i5–1240P @1.7 GHz, the memory is 8.0 GB RAM. This section explores the experimental results obtained by various basic as well as ensemble classifiers. Table 4 demonstrates the performance of classification algorithms in terms of average Accuracy (ACC), Precision, Recall, AUC and F-score respectively. Each dataset's best result is indicated in bold.

Tab. 4. Experimental Results of Classifiers on (a) Arrhythmia, (b) Diabetes Health Indicators, (c) Ionosphere, (d) Sonar, (e) Tic-tac-toe, (f) Waveform datasets respectively

Arrhythmia Dataset					
Classifier	**ACC**	**Prec.**	**Re-call**	**AUC**	**F-Score**
DT	0.72801	0.78774	0.61333	0.72402	0.72641
NB	0.67483	0.75084	**0.64**	0.65159	0.64378
SVM	0.73908	0.86764	0.5	0.71964	0.71810
RF	**0.81637**	0.79232	**0.64**	**0.81199**	**0.81454**
NN	0.68647	0.86764	0.5	0.68457	0.68075
DT+NB+NN+SVM+RF	0.78763	0.90322	0.5	0.77571	0.78088
NN+SVM+RF+DT	0.78995	0.79012	0.62666	0.78668	0.78747
DT+SVM+RF	0.80091	**0.91666**	0.5	0.79206	0.79714

(a)

Diabetes Health Indicators Dataset					
Classifier	**ACC**	**Prec.**	**Re-call**	**AUC**	**F-Score**
DT	0.47376	0.67853	0.59490	0.47375	0.45627
NB	0.71381	**0.69837**	0.59952	0.71377	0.70866
SVM	**0.72583**	0.68734	0.64629	**0.72575**	**0.71142**
RF	0.56773	0.68142	0.64903	0.56771	0.53350
NN	0.66524	0.68290	**0.65497**	0.66520	0.63979
DT+NB+NN+SVM+RF	0.66153	0.68506	0.64856	0.66150	0.63920
NN+SVM+RF+DT	0.62329	0.68546	0.64573	0.62327	0.59322
DT+SVM+RF	0.59304	0.68349	0.64733	0.59302	0.56173

(b)

Ionosphere Dataset					
Classifier	**ACC**	**Prec.**	**Re-call**	**AUC**	**F-Score**
DT	0.90317	0.76017	0.58974	0.88492	0.90092
NB	0.88031	**0.79047**	0.53846	0.84733	0.87489
SVM	0.93444	0.76017	0.58974	0.91351	0.93225
RF	0.93436	0.76269	**0.61538**	0.91987	0.93306
NN	0.92603	**0.79047**	**0.61538**	0.90009	0.92320
DT+NB+NN+SVM+RF	0.93166	0.76017	0.58974	0.91165	0.92964
NN+SVM+RF+DT	0.92880	0.76269	**0.61538**	0.90518	0.92636
DT+SVM+RF	**0.94015**	0.76269	**0.61538**	**0.92486**	**0.93899**

(c)

Sonar Dataset					
Classifier	**ACC**	**Prec.**	**Re-call**	**AUC**	**F-Score**
DT	0.62023	0.76666	0.60606	0.61873	0.61008
NB	0.60714	0.62777	0.36363	0.61602	0.58442
SVM	0.63952	0.74743	0.60606	0.63449	0.62445
RF	**0.73595**	**0.85**	**0.63636**	**0.73072**	**0.72070**
NN	0.64047	0.76666	0.60606	0.63931	0.63070
DT+NB+NN+SVM+RF	0.66880	0.77307	**0.63636**	0.66600	0.64993
NN+SVM+RF+DT	0.66928	0.77307	**0.63636**	0.66827	0.65186
DT+SVM+RF	0.70738	0.79444	**0.63636**	0.70279	0.69524

(d)

Tic-tac-toe Dataset					
Classifier	**ACC**	**Prec.**	**Re-call**	**AUC**	**F-Score**
DT	0.88418	0.73801	0.59595	0.86881	0.88379
NB	0.71714	**0.78245**	0.42424	0.59824	0.65364
SVM	0.89769	**0.78245**	0.54545	0.85634	0.89270
RF	**0.94782**	**0.78245**	**0.61616**	**0.92682**	**0.94648**
NN	0.83820	0.71835	0.54545	0.79607	0.83206
DT+NB+NN+SVM+RF	0.88833	**0.78245**	0.55555	0.84359	0.88240
NN+SVM+RF+DT	0.87995	**0.78245**	0.52525	0.82935	0.87182
DT+SVM+RF	0.93317	**0.78245**	0.57575	0.90695	0.93.116

(e)

Waveform Dataset					
Classifier	**ACC**	**Prec.**	**Re-call**	**AUC**	**F-Score**
DT	0.8116	0.68058	0.56862	0.78975	0.81158
NB	0.8556	0.67619	**0.64117**	0.87362	0.85892
SVM	**0.9038**	0.72810	0.60980	**0.88449**	**0.90296**
RF	0.8892	**0.73584**	0.59019	0.86004	0.88704
NN	0.8728	0.70146	0.61960	0.85659	0.87255
DT+NB+NN+SVM+RF	0.8956	0.72841	0.61176	0.87961	0.89516
NN+SVM+RF+DT	0.8912	0.73021	0.59019	0.85981	0.88867
DT+SVM+RF	0.8948	0.72968	0.59803	0.86844	0.89306

(f)

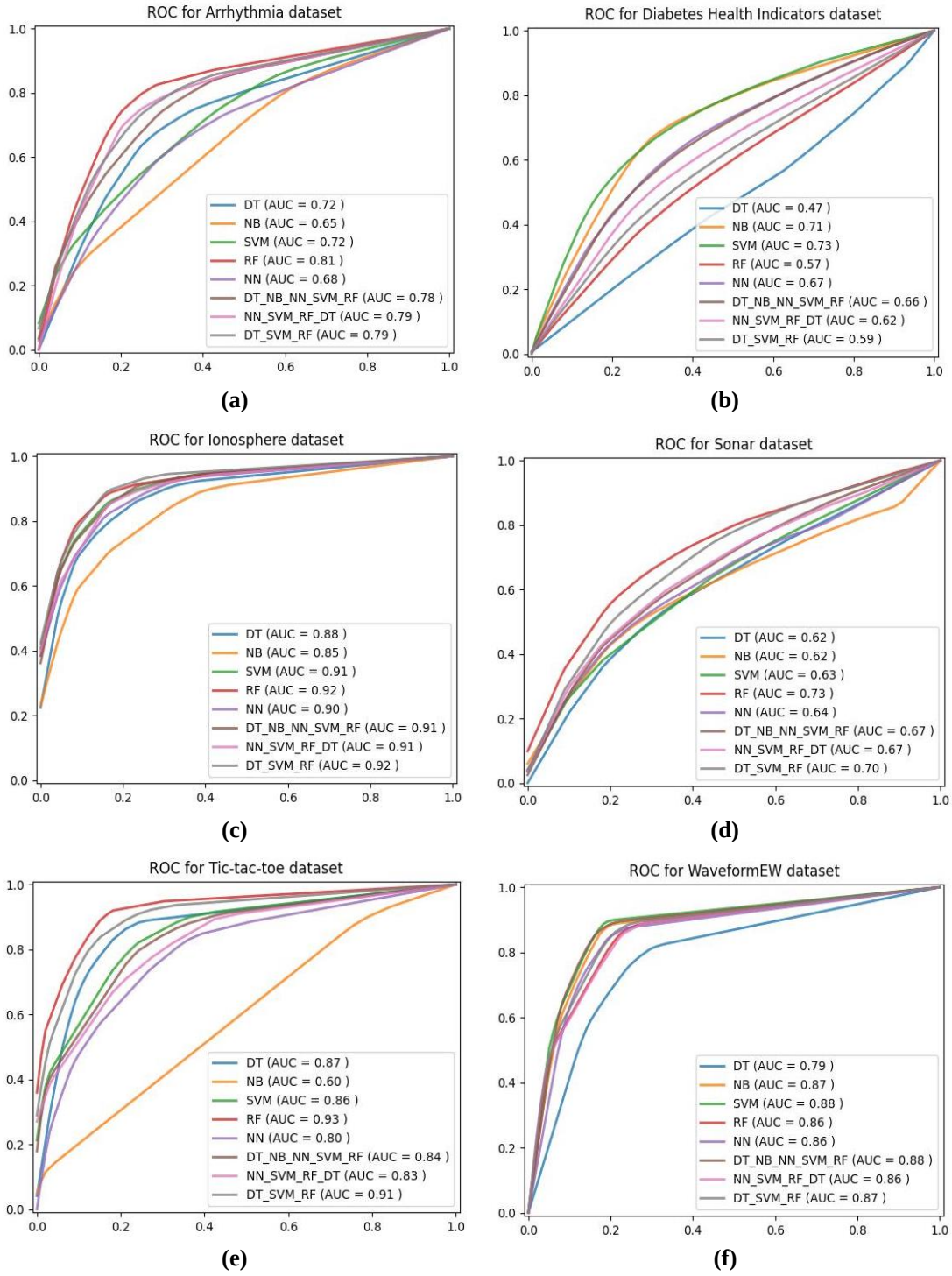

Fig. 7. ROC Curves for (a) Arrhythmia, (b) Diabetes Health Indicators, (c) Ionosphere, (d) Sonar, (e) Tic-tac-Toe, (f) Waveform datasets respectively

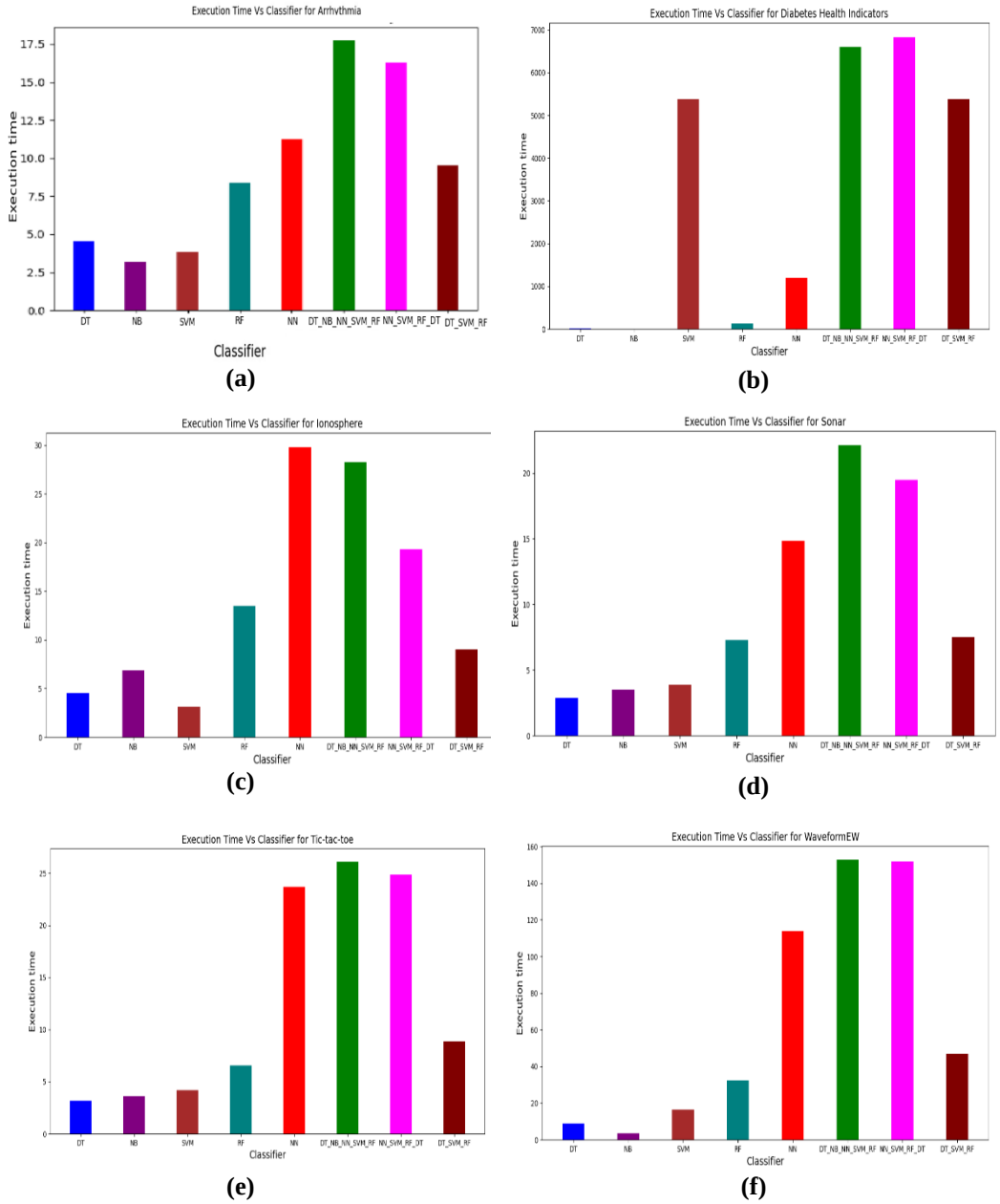

Fig. 8. Bar Graphs with Execution time for each classifier for (a) Arrhythmia, (b) Diabetes Health Indicators, (c) Ionosphere, (d) Sonar, (e) Tic-tac-toe, (f) Waveform datasets

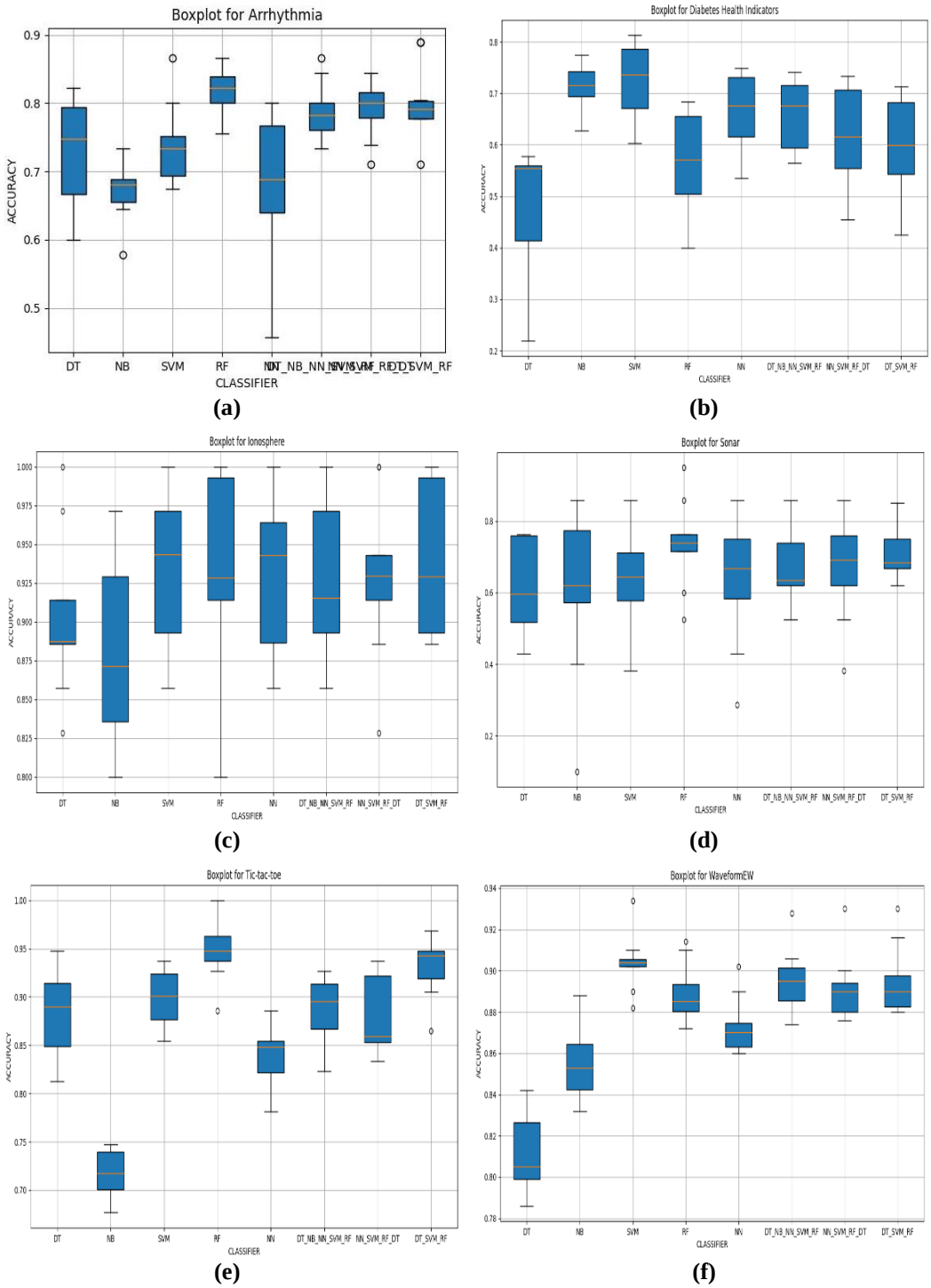

Fig. 9. Box Plots for 10-fold Accuracies of each classifier for (a) Arrhythmia, (b) Diabetes Health Indicators, (c) Ionosphere, (d) Sonar, (e) Tic-tac-toe, (f) Waveform datasets

Precision and recall may vary depending upon the class of interest (i.e., on which class we are considering as positive class) higher precision value indicates less false positives and higher recall indicates less false negatives. The harmonic mean of precision and recall values is F-Score which results low when either precision or recall values are low.

The Arrhythmia dataset achieved highest Accuracy, Recall and F-Score of 81.63%, 64% and 81.45% respectively with RF and the second highest is achieved by the ensemble combination of DT+SVM+RF classifier with accuracy of 80.09% which is very close to the highest accuracy and a highest precision of 91.66% which shows the robust nature of homogeneous ensemble algorithm RF and that the ensemble methods are better in performance when compared with remaining base classifiers. NB has shown an accuracy of 67.48% which is very poor in performance when compared to other classifiers, as Arrhythmia is a high dimensional dataset with 279 attributes and NB assumes that the attributes are independent which is practically not possible. Similarly the second highest precision values is obtained for the ensemble combination of DT+NB+NN+SVM+RF with 90.32% and the lowest precision is again NB classifier and the highest recall value of 75.08% was recorded with NB and also it is observed that the dataset obtained better AUC for RF 81.19% and ensemble combination of classifiers other than the basic classifiers which can be seen in Table 4 (a) and the ROC curve for this dataset with different classifiers along with their respective AUC values were plotted in Fig. 7(a). The execution time can also be considered as a measure for performance evaluation of a classifier and hence the training time of each classifier with arrhythmia dataset is plotted in Fig. 8(a) in which the lowest execution time was given by NB whereas the maximum time was taken by DT+NB+NN+SVM+RF. Boxplot represents the accuracy values of 10-fold cross validation obtained with each classifier and is observed that from Fig. 9(a) that DT+SVM+RF and NN classifiers had given a lesser range and symmetric nature of accuracy scores.

The dataset Diabetes Health Indicators considered is a binary class balanced dataset. Among all the classifiers used, SVM shows better accuracy and an F-score of 72.58% and 71.14% respectively. Similarly, NB gives the second highest score of accuracy, F-score and AUC of 71.38%, 70.86% and 72.57% respectively. Also, the highest precision 69.83% whereas SVM has given the second highest precision score of 68.73%. The Highest recall was given by the robust algorithm NN with 65.49%. The least performance for this dataset was given by DT in all the metrics with an accuracy of 47.37%, Precision of 67.85%, recall of 59.49%, AUC of 47.37% and F-Score of 45.62% which can be seen in Table 4 (b) and ROC curve for the same AUC can be observed in Fig. 7(b). This dataset has lowest training time with NB and then DT but DT+NB+NN+SVM+RF has the highest execution time compared to the remaining classifiers, as can be observed in Fig. 8(b). Boxplot for accuracy scores of 10-fold cross validation are been given in Fig. 9(b) in which the lesser range of values are indicated by NB when compared to other classifiers.

The Ionosphere dataset has achieved the best performance with the ensemble combination of DT+SVM+RF with an accuracy of 94.01%, F-Score of 93.89% and recall of 61.53% and AUC of 92.48%. However, it is observed that the same recall score as of highest value is also given by RF, NN and the ensemble combination of NN+SVM+RF+DT. The lowest accuracy and recall were recorded by NB, precision by DT, SVM and the ensemble combination of DT+NB+NN+SVM+RF which can be seen in Table 4(c). ROC curves with their corresponding AUC were shown in Fig. 7(c) and the maximum and minimum AUC were given by NN and NB respectively. Fig. 8(c) gives the bar graph of execution time in

which the minimum execution time was taken by RF and the maximum time was taken by SVM. Boxplot representation for the accuracy scores obtained from 10-fold cross validation can be seen in Fig. 9(c) and can be observed that almost all the classifiers had given a wider range of accuracy scores and NN+SVM+RF+DT had given a lesser range of scores.

The Sonar was also a binary class classification dataset and consists of high dimensional data with 60 attributes. From Table 4(d) we can observe that RF had given a better performance with regard to all the metrics such as accuracy, precision, recall, f-score of 73.59%, 85%, 63.63% and 72. 07 % respectively and a maximum AUC is of 73.07% is also obtained with RF which can be seen in the ROC plot in Fig. 7(d) when compared to all other classifiers and Similarly NB has given the least performance with respect to all the metrics for the sonar dataset. The Ensemble combination DT+SVM+RF has scored the second highest accuracy oy 70.73% and also all the ensemble combinations has achieved a recall score of 63.63% as of the highest score. NN has given a lesser range of accuracy scores with the 10-fold cross validation and also similar group of lesser range values are given by the ensemble combination DT+SVM+RF which can be observed in the boxplot plotted in Fig. 9(d). Also, the execution time of Sonar dataset with different classifiers were shown in Fig. 8(d) in which DT has taken minimum time and the Ensemble combination of DT+NB+NN+SVM+RF has taken the maximum execution time.

The ensemble classifier RF has given highest performance with the dataset Tic-tac-toe with accuracy of 94.78%, precision of 78.24%, Recall of 61.61%, and F-score of 94.64% and also the same precision value as of the highest precision were given by SVM and all the ensemble combinations. Similarly, the second highest accuracy was given by DT+SVM+RF with the score of 93.31% which can be observed in Table 4(e). Fig. 8(e) shows the execution time of each classifier with Tic-tac-toe dataset, in which DT has taken minimum time and the Ensemble combination of DT+NB+NN+SVM+RF has taken the maximum execution time. The Box plot for 10 – fold cross validation accuracy scores was plotted and can be seen in Fig. 9(e) in which the lesser range of values with the highest scores were given by NN and the ensemble combination DT+SVM+RF which is very close to NN. Fig. 7(e) shows the ROC plot of each classifier and their corresponding AUC values.

The dataset Waveform is a multiclass Classification dataset with 3 classes and SVM with an accuracy of 90.38% and F-score of 90.29% and the second highest accuracy was given by the ensemble combination DT+NB+NN+SVM+RF and the highest precision was given by RF with 73.58% and recall was given by NB with 64.11% and also from Table 4(f) it can be observed that all the ensemble combinations were very close to the performance of the classifier which has the highest accuracy From Fig. 7(f) it can be observed that SVM has achieved the highest AUC of 88.44% and the second highest AUC was given by NB with 87.36%. Similarly, the Boxplot for Waveform Dataset was shown in Fig. 9(f) which shows again that SVM has given the lesser range of accuracy scores and the highest value when compared to remaining algorithms. The minimum execution time was taken by NB and Maximum time was taken by DT+NB+NN+SVM+RF which can be observed from Fig. 8(f).

Tab. 5. Comparison of performance of the proposed method with existing works

S. No.	Dataset	Reference	Methodology	Accuracy in (%)
1	Arrhythmia	(Ecemiş et al., 2022)	RBF	60.00
			MLP+SVM+RBF+RF	71.00
			MLP+RF	66.00
			RF1+RF2	72.00
			RF1+(RF2+MLP+SVM+RBF)	71.00
		(Gupta et al., 2014)	SVM -Poly degree2	66.00
			RF+SVM	77.40
			Pattern	69.00
			Net Two-level RF	70.00
		(Yogita et al., 2020)	RF+PCA	71.00
			SVM+PCA	73.00
			SVM+PCA with feature selection	65.00
			Kernalized SVM	71.00
			Kernalized SVM +PCA	75.00
			Kernalized SVM + Feature Selection Kbest	77.00
		(bin Basir & binti Ahmad, 2017)	RIPPER	73.67
			Boosting +RIPPER	73.41
			Bagging+RIPPER	73.80
			PART	74.13
			Boosting +PART	74.98
			Bagging+PART	76.93
			PRISM	62.36
			Boosting +PRISM	61.71
			Bagging+PRISM	66.07
			OneR	59.76
			Boosting +OneR	59.69
			Bagging+OneR	59.37
		(Shi et al., 2022)	edRVFL	72.43
			WedRVFL	73.22
			PedRVFL	73.66
			WpedRVFL	73.88
		Proposed Method	DT+SVM+RF	80.9
2	Ionosphere	(bin Basir & binti Ahmad, 2017)	RIPPER	92.87
			Boosting +RIPPER	93.63
			Bagging+RIPPER	93.54
			PART	90.78
			Boosting +PART	92.62
			Bagging+PART	91.95
			PRISM	89.77
			Boosting +PRISM	91.87
			Bagging+PRISM	91.03
			OneR	87.26
			Boosting +OneR	91.53
			Bagging+OneR	87.17
		(Ngo et al., 2022)	Bagging	94.00
			Extra Trees	94.30
		Proposed Method	DT+SVM+RF	94.02

Tab. 5. Comparison of performance… – cont.

S. No.	Dataset	Reference	Methodology	Accuracy in (%)
3	Waveform	(Shi et al., 2022)	edRVFL WedRVFL PedRVFL WpedRVFL	86.17 86.92 86.98 87.13
		(Alshdaifat et al., 2021)	Majority voting with (NB+DT+RB+KNN+ANN+SVM) Majority voting with Optimal Classifier Selection(OCS)	89.00 89.00
		Proposed Method	DT+NB+NN+SVM+RF NN+SVM+RF+DT DT+SVM+RF	89.56 89.12 89.48
4	Tic-tac-toe	(Hongle et al., 2022)	CSEL Balance Cascade Easy Ensemble Ada Boost	80.58 78.08 86.98 79.75
		Proposed Method	DT+SVM+RF	93.30

When considering the Proposed Assembly Combinations with existing methods, it can be seen that the authors' proposed combination of assemblies gave the best performance. In (Ecemiş et al., 2022) the authors constructed the nested classifiers using the base classifiers like MLP, Radial Bias Function (RBF), SVM, RF out of which the random forest classifier combination RF1+RF2 has given good performance in predicting the cardiac arrhythmia. The authors (Gupta et al., 2014) has implemented a classifier with linear kernel SVM and RF which gave a generalization error of 77.4%. Feature selection along with kernelized SVM classifier is applied (Yogita et al., 2020) and (bin Basir & binti Ahmad, 2017; Shi et al., 2022). The authors compared different combination of ensemble classifiers for optimizing the performance of cardiac arrhythmia, waveform and Ionosphere (Ngo et al., 2022) datasets along with few more datasets. The authors (Hongle et al., 2022) used minimum redundancy and maximum correlation to select base classifiers and showed that clustering under sample (CSEL) has better accuracy, of which our model outperforms conventional methods proposed in previous work, as can be observed in Table 5.

7. CONCLUSION AND FUTURE SCOPE

In this paper, heterogeneous ensemble classifiers are proposed and their performance is evaluated on various datasets. Performance of ensemble classifiers is also compared with basic classifiers. The results of comparison analysis are presented in this paper, which used five of the most well-known basic ML algorithms for classification, including DT, NN, NB, SVM and RF along with their ensemble combinations DT+NB+NN+SVM+RF, NN+SVM+RF+DT and DT+SVM+RF on five various publicly available UCI repository datasets and one Diabetes health Indicators dataset from kaggle repository. The highest accuracy of 94.01% is achieved by the ensemble classifier (DT+SVM+RF) for ionosphere dataset. Basic classifier SVM has achieved highest accuracy of 72.58% with Diabetes Health Indicators dataset and 90.38% accuracy with Waveform datasets. The accuracy of RF for

Arrhythmia, Sonar and Tic-tac-toe datasets are 81.63%, 73.59%, and 94.78% respectively. From the experimental results it is observed that in most of the cases either the homogeneous ensemble classifier or the combination of basic classifiers i.e., heterogeneous ensemble has given better results than the basic classifiers. It is difficult to ensemble heterogeneous classifiers as it is complex and requires more resources to train and maintain the model especially with the large datasets. If the training dataset is small or biased, the ensemble may not perform well on unseen data. This study does not focus on optimizing the hyperparameters of the classifiers used.

It is observed that the proposed ensemble combination outperformed some of the existing models. Finally, it can be concluded that each classifier has its own set of benefits and limitations.

In future, the proposed heterogeneous ensemble classifiers can again be ensembled with some other classifiers along with some feature selection algorithms and hyperparameter optimization to improve the classification performance.

REFERENCES

Alshayeji, M. H., Ellethy, H., Abed, S., & Gupta, R. (2022). Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach. *Biomedical Signal Processing and Control*, *71*(PA), 103141. https://doi.org/10.1016/j.bspc.2021.103141

Alshdaifat, E., Al-hassan, M., & Aloqaily, A. (2021). Effective heterogeneous ensemble classification: An alternative approach for selecting base classifiers. *ICT Express*, *7*(3), 342–349. https://doi.org/10.1016/j.icte.2020.11.005

Baumann, P., Hochbaum, D. S., & Yang, Y. T. (2019). A comparative study of the leading machine learning techniques and two new optimization algorithms. *European Journal of Operational Research*, *272*(3), 1041–1057. https://doi.org/10.1016/j.ejor.2018.07.009

bin Basir, M. A., & binti Ahmad, F. (2017). New Feature Selection Model Based Ensemble Rule Classifiers Method for Dataset Classification. *International Journal of Artificial Intelligence & Applications*, *8*(2), 37–43. https://doi.org/10.5121/ijaia.2017.8204

Chandrika, Divya, C., Gowramma, G. S., & Varun, C. R. (2018). A comparative analysis on evaluation of classification algorithms based on ionospheric data. *International Journal of Computer Sciences and Engineering*, *6*(5), 636–640. https://doi.org/10.26438/ijcse/v6i5.636640

Consuegra-Ayala, J. P., Gutiérrez, Y., Almeida-Cruz, Y., & Palomar, M. (2022). Intelligent ensembling of auto-ML system outputs for solving classification problems. *Information Sciences*, *609*, 766–780. https://doi.org/10.1016/j.ins.2022.07.061

Ecemis, C., Acu, N., & Sari, Z. (2022). Classification of Imbalanced Cardiac Arrhythmia Data. *European Journal of Science and Technology*, *34*, 546-552. https://doi.org/10.31590/ejosat.1083423

Fang, X., Klawohn, J., De Sabatino, A., Kundnani, H., Ryan, J., Yu, W., & Hajcak, G. (2022). Accurate classification of depression through optimized machine learning models on high-dimensional noisy data. *Biomedical Signal Processing and Control*, *71*(Part B), 103237. https://doi.org/10.1016/j.bspc.2021.103237

Farhat, N. H. (1992). Photonit neural networks and learning mathines the role of electron-trapping materials. *IEEE Expert-Intelligent Systems and Their Applications*, *7*(5), 63–72. https://doi.org/10.1109/64.163674

Fath, A. H., Madanifar, F., & Abbasi, M. (2020). Implementation of multilayer perceptron (MLP) and radial basis function (RBF) neural networks to predict solution gas-oil ratio of crude oil systems. *Petroleum*, *6*(1), 80–91. https://doi.org/10.1016/j.petlm.2018.12.002

Ganie, S. M., & Malik, M. B. (2022). An Ensemble Machine Learning Approach for Predicting Type-II Diabetes Mellitus based on Lifestyle Indicators. *Healthcare Analytics*, *2*, 100092. https://doi.org/10.1016/j.health.2022.100092

Gupta, V., Srinivasan, S., & Kudli, S. S. (2014). *Prediction and Classification of Cardiac Arrhythmia*. https://cs229.stanford.edu/proj2014/Vasu%20Gupta,%20Sharan%20Srinivasan,%20Sneha%20Kudli,%20Prediction%20and%20Classification%20of%20Cardiac%20Arrhythmia.pdf

Hongle, D., Yan, Z., Lin, Z., Yeh-Cheng, C., Gang, K., & Chen, Y.-C. (2022). Selective Ensemble Learning Algorithm for Imbalanced Dataset. *Preprint*. https://doi.org/10.21203/rs.3.rs-721493/v1

Jia, J., & Qiu, W. (2020). Research on an ensemble classification algorithm based on differential privacy. *IEEE Access*, *8*, 93499–93513. https://doi.org/10.1109/ACCESS.2020.2995058

Kilincer, I. F., Ertam, F., & Sengur, A. (2021). Machine learning methods for cyber security intrusion detection: Datasets and comparative study. *Computer Networks*, *188*, 107840. https://doi.org/10.1016/j.comnet.2021.107840

Kushwah, J. S., Kumar, A., Patel, S., Soni, R., Gawande, A., & Gupta, S. (2021). Comparative study of regressor and classifier with decision tree using modern tools. *Materials Today: Proceedings*, *56*(6), 3571-3576. https://doi.org/10.1016/j.matpr.2021.11.635

Ma, T. M., Yamamori, K., & Thida, A. (2020). A comparative approach to naïve bayes classifier and support vector machine for email spam classification. *2020 IEEE 9th Global Conference on Consumer Electronics, GCCE 2020* (pp. 324–326). IEEE. https://doi.org/10.1109/GCCE50665.2020.9291921

Maniruzzaman, M., Jahanur Rahman, M., Ahammed, B., Abedin, M. M., Suri, H. S., Biswas, M., El-Baz, A., Bangeas, P., Tsoulfas, G., & Suri, J. S. (2019). Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms. *Computer Methods and Programs in Biomedicine*, *176*, 173–193. https://doi.org/10.1016/j.cmpb.2019.04.008

Mohamed, A. R. (2017). Comparative Study of Four Supervised Machine Learning Techniques for Classification. *International Journal of Applied Science and Technology*, *7*(2), 5–18.

Nazari, E., Aghemiri, M., Avan, A., Mehrabian, A., & Tabesh, H. (2021). Machine learning approaches for classification of colorectal cancer with and without feature selection method on microarray data. *Gene Reports*, *25*, 101419. https://doi.org/10.1016/j.genrep.2021.101419

Ngo, G., Beard, R., & Chandra, R. (2022). Evolutionary bagging for ensemble learning. *Neurocomputing*, *510*, 1-14. https://doi.org/10.1016/j.neucom.2022.08.055

Patel, H. H., & Prajapati, P. (2018). Study and analysis of decision tree based classification algorithms. *International Journal of Computer Sciences and Engineering*, *6*(10), 74–78. https://doi.org/10.26438/ijcse/v6i10.7478

Patel, N., & Upadhyay, S. (2012). Study of various decision tree pruning methods with their empirical comparison in WEKA. *International Journal of Computer Applications*, *60*(12), 20–25. https://doi.org/10.5120/9744-4304

Priyanka, & Kumar, D. (2020). Decision tree classifier: A detailed survey. *International Journal of Information and Decision Sciences*, *12*(3), 246–269. https://doi.org/10.1504/ijids.2020.108141

Pugliese, R., Regondi, S., & Marini, R. (2021). Machine learning-based approach: global trends, research directions, and regulatory standpoints. *Data Science and Management*, *4*, 19–29. https://doi.org/10.1016/j.dsm.2021.12.002

Punyapornwithaya, V., Klaharn, K., Arjkumpa, O., & Sansamur, C. (2022). Exploring the predictive capability of machine learning models in identifying foot and mouth disease outbreak occurrences in cattle farms in an endemic setting of Thailand. *Preventive Veterinary Medicine*, *207*, 105706. https://doi.org/10.1016/J.PREVETMED.2022.105706

Qian, X., Zhou, Z., Hu, J., Zhu, J., Huang, H., & Dai, Y. (2021). A comparative study of kernel-based vector machines with probabilistic outputs for medical diagnosis. *Biocybernetics and Biomedical Engineering*, *41*(4), 1486–1504. https://doi.org/10.1016/j.bbe.2021.09.003

Revathi, A., Kaladevi, R., Ramana, K., Jhaveri, R. H., Kumar, M. R., & Kumar, M. S. P. (2022). Early detection of cognitive decline using machine learning algorithm and cognitive ability test. *Security and Communication Networks*, *2022*, 4190023. https://doi.org/10.1155/2022/4190023

Rezvani, S., & Wang, X. (2022). Neurocomputing intuitionistic fuzzy twin support vector machines for imbalanced data. *Neurocomputing*, *507*, 16–25. https://doi.org/10.1016/j.neucom.2022.07.083

Sevinç, E. (2022). An empowered AdaBoost algorithm implementation: A COVID-19 dataset study. *Computers and Industrial Engineering*, *165*, 107912. https://doi.org/10.1016/j.cie.2021.107912

Shafi, A. S. M., Molla, M. M. I., Jui, J. J., & Rahman, M. M. (2020). Detection of colon cancer based on microarray dataset using machine learning as a feature selection and classification techniques. *SN Applied Sciences*, *2*(7), 1–8. https://doi.org/10.1007/s42452-020-3051-2

Shi, Q., Suganthan, P. N., & Katuwal, R. (2022). Weighting and pruning based ensemble deep random vector functional link network for tabular data classification. *arXiv:2201.05809*. http://arxiv.org/abs/2201.05809

Swathy, M., & Saruladha, K. (2021). A comparative study of classification and prediction of cardio-vascular diseases (cvd) using machine learning and deep learning techniques. *ICT Express*, *8*(1), 109-116. https://doi.org/10.1016/j.icte.2021.08.021

Tewari, S., & Dwivedi, U. D. (2020). A comparative study of heterogeneous ensemble methods for the identification of geological lithofacies. *Journal of Petroleum Exploration and Production Technology*, *10*(5), 1849–1868. https://doi.org/10.1007/s13202-020-00839-y

Thirunavukkarasu, K., Singh, A. S., Rai, P., & Gupta, S. (2018). Classification of IRIS dataset using classification based KNN Algorithm in supervised learning. *2018 4th International Conference on Computing Communication and Automation, ICCCA 2018* (pp. 4–7). IEEE. https://doi.org/10.1109/CCAA.2018.8777643

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1), 1–16. https://doi.org/10.1186/s12911-019-1004-8

Wade, B. S. C., Joshi, S. H., Gutman, B. A., & Thompson, P. M. (2017). Machine learning on high dimensional shape data from subcortical brain surfaces: A comparison of feature selection and classification methods. *Pattern Recognition*, *63*, 731–739. https://doi.org/10.1016/j.patcog.2016.09.034

Wei, X., Zou, N., Zeng, L., & Pei, Z. (2022). PolyJet 3D printing: Predicting color by multilayer perceptron neural network. *Annals of 3D Printed Medicine*, *5*, 100049. https://doi.org/10.1016/j.stlm.2022.100049

Yakut, Ö., & Bolat, E. D. (2022). A high-performance arrhythmic heartbeat classification using ensemble learning method and PSD based feature extraction approach. *Biocybernetics and Biomedical Engineering*, *42*(2), 667–680. https://doi.org/10.1016/j.bbe.2022.05.004

Yogita, B., Akanksha, M., Shefali, A., Tanya, M., & Gresha, B. (2020). Classification of Cardiac Arrhythmia Using Kernelized SVM. *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)* (pp. 922-926). IEEE. https://doi.org/10.1109/ICOEI48184.2020.9143000.