Published in AJSE, Vol:23, Issue: 1
Received on 23rd August 2023
Revised on 27th January 2024
Published on 25th April 2024

AIUB JOURNAL OF SCIENCE AND ENGINEERING
ISSN: 1608 – 3679 (print) 2520 – 4890 (Online)
DOI: 10.53799/ajse.v23i1.877

WVEHDD: Weighted Voting based Ensemble System for Heart Disease Detection

Usha Rani Gogoi*

***Abstract*— Although several machine learning (ML) based algorithms are proposed by various researchers for heart disease detection (HDD), most of these works considered a very small experimental dataset to justify the efficiency of ML techniques in HDD. Moreover, despite of the low correlation of the features with the target, all the features were used for HDD. Considering the limitations of these existing systems, current study emphasizes on the designing of a Weighted Voting based Ensemble (WVE) Classifier for HDD from a sufficiently large dataset comprising of 1296 instances. Although there are 13 features, only 4 features are found to be statistically significant in HDD. For designing an efficient WVE classifier for HDD, the weighted votes of five efficient classifiers are combined to get the final decision. The experimental result shows that the proposed WVEHDD system outperforms the existing systems by providing the highest train accuracy of 96.15% and test accuracy of 95.64%.**

***Index Terms*— Heart disease, Feature selection, Machine learning, Weighted voting-based ensemble classifier, Classification;**

I. Introduction

THE leading cause of death worldwide, accounting for 17.9 million fatalities, or 31% of all deaths, is cardiovascular disease (CVD) [1] . Out of every five deaths, strokes and heart attacks account for more than four deaths [1]. According to World Health Organization, unhealthy eating habits, low physical activity, alcohol abuse, and tobacco use are some of the potential risk factors that accelerate heart-related complications like high blood pressure, high blood glucose levels, extreme blood lipids, overweight, and obesity etc. [2]. One of the highest burdens of CVD worldwide is seen in India. In India, the annual death from CVD was projected to rise from 2.26 million in 1990 to 4.77 million in 2020 [3]. This increasing burden of heart disease (HD) patients incurs the need for a system that is efficient and affordable for providing a preliminary assessment of a patient based on the results of his or her medical tests. Doctors use several indicators to identify and diagnose HD. But considering the growing population and the number of HD patients, evaluating the medical records manually is a highly difficult and time-consuming process and sometimes leads to imprecise diagnoses. However, with the emergence of machine learning (ML), it may now be used in the health industry to diagnose, detect, and forecast various ailments in a non-invasive way. In the literature, it has been noted that many ML-based HD diagnosis systems were developed by different researchers employing various HD disease datasets. One of the most used HD datasets is the CHD dataset [4]. The quantity of the training dataset affects how well the ML approach performs. If balanced datasets are utilized for model training and testing, ML models perform better [5, 6]. Furthermore, by incorporating appropriate and pertinent features from the data, the model's prediction abilities may be improved. Therefore, data balancing and feature selection are crucial to increase the model performance. As reported in Section II, although different diagnosis methods have been presented by various researchers in the literature; almost all these methods used small experimental datasets. Moreover, most of the research works considered all 13 features given on the dataset. On the contrary, this work emphasizes the identification of the risk factors of HD and evaluates the performance of ten different traditional ML techniques including Logistic Regression (LR), Support vector machines (SVM), K-nearest neighborhood (KNN), Naïve Bayes (NB), Artificial neural network (ANN), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GBoost), Extreme gradient boost (XGBoost) and Stochastic gradient descent (StGD) in HD prediction. The classification performance of each of these algorithms are evaluated by using various evaluation metrics such as classification accuracy, sensitivity, specificity, precision, F1-score and Area Under the Receiver Operating Characteristic Curve (AUC) score. Then by considering the five most efficient ML models, a weighted voting-based ensemble (WVE) classifier is designed for HDD The result of the hybrid classifier is compared with the other state-of-the-art classification algorithms which proves that our hybrid classifier outperforms the other state-of-the art ML models using only 4 HD parameters. The overview of the proposed WVEHDD System is demonstrated in Figure 1.

The major contributions of this paper are as follows:

1) To determine the efficient classification algorithms that would efficiently classify the HD data.

2) Unlike the other works, this work focuses on finding the most discriminative features, and their efficiency in HD

Usha Rani Gogoi is working as Assistant Professor in the Department of Computer Science and Engineering, The Neotia University (TNU), Diamond Harbour, Kolkata, West Bengal, India (Email: ushagogoi.cse@gmail.com)

prediction is evaluated by using traditional as well as ensemble ML techniques.

3) The efficiency of the classification models is evaluated by using more than one state-of-the-art evaluation metric. Both the training as well as testing accuracies are demonstrated in the paper.

4) To have an unbiased decision on a classification model, the classification performances are compared with other relevant existing systems.

The remainder of the paper is structured as follows. Section II describes the related works along with their advantages and limitations. Section III presents the methodology utilized related to data preparation, data analysis, feature selection and ML models training testing and ML models' performance evaluation. In Section IV, the discussion of the analysis and findings are made. Section V concludes the paper.

II. Literature Survey

In the literature, researchers suggested different ML-based methods to diagnose HD. In order to highlight the significance of the proposed work, this study presents some existing ML-based diagnosis techniques. In [7], S. Kumar et al. proposed a HD prediction system, where they used C4.5 in combination with genetic algorithm (GA) to detect HDs. They reported the highest training accuracy of 74.82% and testing accuracy of 73.20% on UCI dataset [8]. Their proposed C4.5 algorithm generated 7 rule for HDD.

In [9], Bashir et al. developed an ensemble-based model that combined NB, Gini Index (GI) based DT, information Gain (IG) based DT, instance-based learner (IBL), and SVM models for the prediction of HD. They obtained an accuracy of 87.37% with their proposed ensemble technique on CHD dataset [4].

K. Polat [10] proposed a new attribute weighting method to classify the samples into normal and abnormal classes. Four classification algorithms: Linear discriminant analysis (LDA), KNN, SVM and RF were used to evaluate the performance of their proposed method. The highest classification accuracy of 96.63% was obtained with RF.

S. A. E. Mienye et al. [11], proposed a novel HDD system that integrated an enhanced sparse encoder (SAE) to a SoftMax regression. They achieved a classification accuracy of 91%.

Dan Gan et al. [12], proposed a HDD system by integrating TANBN with sensitive classification algorithm. Their proposed system can handle imbalanced dataset and obtained

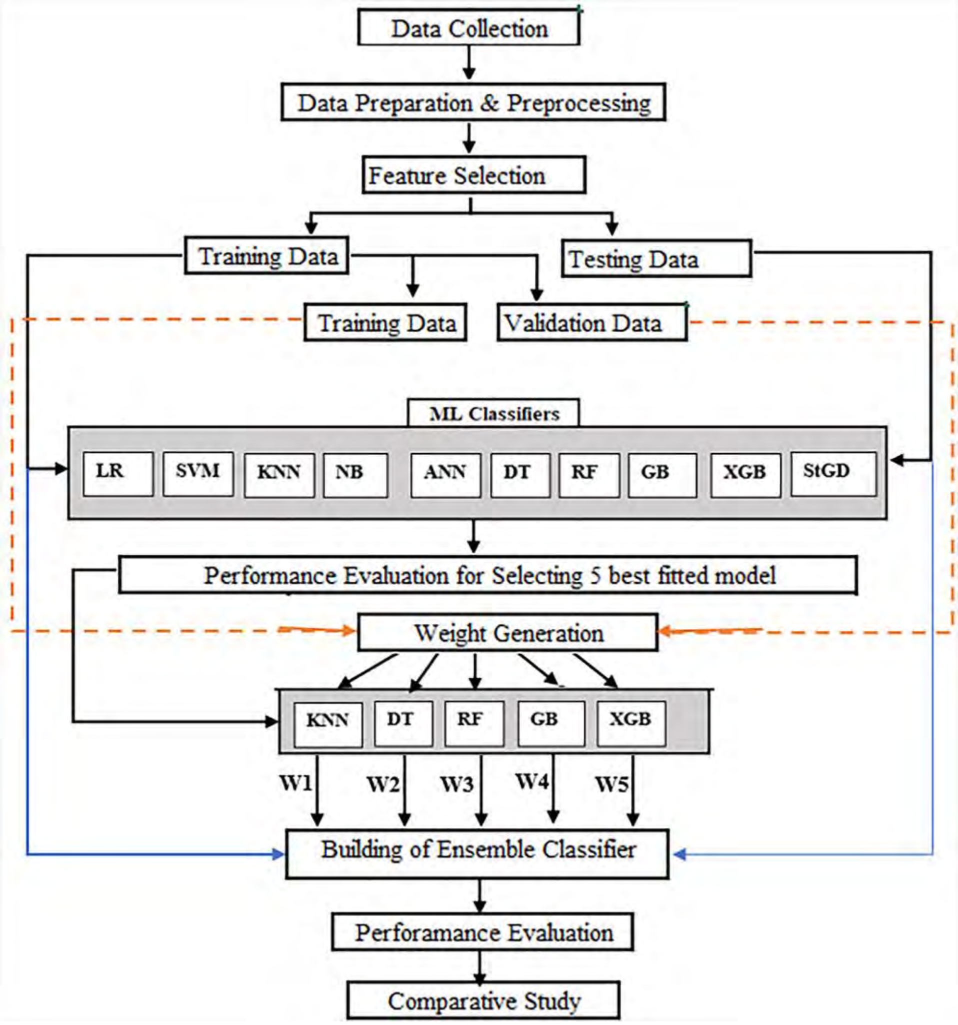

Fig. 1. The workflow of the proposed WVEHDD system

the highest accuracy of 80.27% and AUC of 88.27%.

D. Shah et al. [13] used various ML techniques including NB, DT, KNN, and RF algorithms for prediction of HD. Among all these methods, the highest classification accuracy of 88.16% was obtained with NB with the testing dataset.

In [14], authors evaluated the performance of several ML techniques: KNN, SVM, LR, RF, NB, Ensembled classifier, ^{i}and ANN in HD prediction. They reported that among all these models, the highest classification accuracy of 96.1% was obtained with ANN in UCI dataset consisting of 303 records.

R. Atallah et al. [15] proposed a majority voting ensemble method by combining Stochastic Gradient Descent (SGD), KNN, RF and LR to predict the presence of HD. Their proposed ensemble method provided the highest accuracy of 90% on UCI dataset.

D. K Plati et al. [16] proposed an ML based chronic heart failure diagnosis system. Their proposed method comprises of 3 stages: preprocessing, feature selection and classification. They reported the highest classification accuracy of 91.23% in a dataset with 487 subjects. The summary of all of the above-mentioned existing works is provided in Table I.

III. Proposed Methodology

The details of the research materials and techniques are discussed in the following subsections.

A. Experimental Dataset

The datasets used in this paper are publicly available. To increase the size of the experimental dataset, we combined the instances of Cleveland, Hungary, Switzerland, Long Beach V and Statlog (Heart) databases. The combined experimental dataset comprises of 1296 samples where each of the samples is described by 14 attributes. The details of the attributes are provided in Table II.

B. Pre-processing of the Dataset

Before feeding the data for model building, it is important to transform the raw data to make accurate predictions. However, there is no missing value in the dataset. Although there are outliers in the raw dataset, but the dataset containing only the statistically significant features does not contain any outliers for which the only pre-processing method we performed was normalizing the dataset.

TABLE I
Summary of the Existing ML based heart disease detection (HDD) systems

Authors	ML Methods used	Dataset Details	No of features	Accuracy	Advantages and Disadvantages
S. Kumar et al. [7], 2018	C4.5 DT & GA	CHD: 303	13	73.20 %	The performance of the proposed system is poor.
Bashir et al. [9], 2016	Ensemble approach (NB, DT (GI), DT (IG), IBL, SVM)	Statlog: 270 CHD:303	13 & 13	**Statlog:** Acc - 87.37%, Sens. –87.50%, Spec. –87.27%, F-measure - 87..38% **UCI:** Acc - 81.82%, Sens –73.68%, Spec –92.86%, F-measure - 82.17%	Both the experimental datasets are small and the accuracy of their proposed system on both the datasets are below 90%.
K Polat et al. [10], 2018	DA, KNN, SVM, RF	SPECT: 267	22	Acc - 96.637%, Prec. – 97.10%, Rec –96.60%, AUC – 99.00%, F-measure - 96.70%	The experimental dataset is small.
S. A. E. Mienye et al. [11], 2020	SAE + Softmax	Framingham Heart Study Dataset: 4238 samples	16	Acc - 91%, Prec –93%, Rec –90%, F-measure -92%	The experimental dataset is quite large and the performance is also above 90%. However, they did not perform feature selection.
Dan Gan et al. [12], 2020	AdaCTANBN	CHD: 300	13	Acc – 80.27%, Prec –88.873%,	The experimental dataset is small and classification performance is <90%
D. Shah et al. [13], 2020	NB, DT, KNN and RF	CHD: 303	13	Train Acc – 91.78% (KNN) Test Acc – 88.16% (NB)	The experimental dataset is small.
Muhammad Waqar et al. [14], 2021	KNN, SVM, LR, RF, NB, ANN, Ensemble classifier	UCI: 303	13	Acc – 96.1%, Prec – 95.7% Rec – 95.7%, F1 Score – 96%	The experimental dataset is small.
R. Atallah et al. [15], 2019	Stochastic Gradient Descent (SGD), KNN, RF, LR, Ensemble Classifier by combining all	CHD: 303	13	SGD - 88% KNN - 87% RF - 87% LR - 87% Ensemble - 90%	The experimental dataset is small and their reported highest accuracy is 90%
D. K Plati et al. [16], 2021	RF, Rotational Forest, NB, KNN, SVM, LMT, BN	Private dataset: 487 instances	--	Acc – 91.23% Sens – 93.83% Spec – 89.62	The experimental dataset is small.

*Acc – Accuracy, Spec – Specificity, Sens – Sensitivity, Rec – Recall, Prec – Precision

TABLE II
DETAILS OF THE ATTRIBUTES OF HEART DISEASE DATASET

SL No.	Attributes' Name	Parameter Values & Description	Statistical Significance (Between HD and non-HD group)
1.	age	Age in years	5.10541e+5
2.	sex	Gender (Male: 1 & Female: 0)	5.0408e+5
3.	**cp**	**Chest Pain (0 – Atypical angina, 1- typical angina, 2 – asymptomatic, 3 – non-anginal pain)**	**2.3752e-68**
4.	trestbps	Resting blood pressure (in mmHg (unit))	4.7937e+5
5.	chol	serum cholesterol in mg/dl	4.8025e+5
6.	fbs	Fasting Blood Sugar > 120 mg/dl (1 – true, 0 – false)	4.5107e+5
7.	**restecg**	**Resting ECG: (0 - normal 1 - having ST-T wave abnormality 2 - left ventricular hypertrophy)**	**1.2486e-7**
8.	**thalach**	**Maximum heart rate achieved**	**9.2858e-58**
9.	exang	Exercise induced angina: (1 – yes, 0 – no)	5.3986e+5
10	oldpeak	ST depression induced by exercise relative to rest	5.6258e+5
11	**slope**	**Slope of the peak exercise ST segment: (0 – upsloping, 1 – flat, 2 – downsloping)**	**2.1057e-43**
12	ca	Number of major vessels	5.5517e+5
13	thal	Thalassemia: (0 = normal, 1 = fixed defect, 2 = reversible defect)	5.4007e+5
14	Target	—	

C. *Feature Selection*

The preprocessing of the dataset is followed by the feature selection as in ML, feature selection plays a crucial role. The complexity of the HDD system can be reduced by discarding the unnecessary and redundant features that do not add much value to HDD. As the dataset comprises of a different number of samples in both HD and non-HD groups, so for the selection of the statistically significant features, we have used the Wilcoxon rank sum test [17]. The statistical test of significance finds out only those features that reach the significance of p<0.005 and these features are considered as the most discriminative features for HD detection. The significance level of each feature value is measured against the null hypothesis "there is no significant difference between the non-HD subject and the HD subject" and tabulated in Table II. As illustrated in Table II, out of 13 features (excluding the target), only four features: chest pain (cp), resting ecg result (restecg), maximum heart rate achieved (thalach) and slope are found to be statistically significant in HDD. Like the p-values, as demonstrated in Fig. 2, there is a significant feature value difference between the non-HD and the HD subjects. Fig. 2(a) shows that the subjects from the non-HD group suffer from Atypical angina while, the subjects from the HD group suffer more from Typical angina. Similarly, as illustrated in Fig. 2(b), exercise-induced angina is more probable in subjects of HD group. As illustrated in Fig. 2(c), the maximum heart rate is more in subjects of HD group in comparison to the subjects of non-HD group. Likewise, the presence of abnormality is also prominent in subject of HD group as depicted in Fig. 2(d).

D. *Selection of Efficient ML models*

For designing an efficient ML System for HD prediction, we need to find out the most potential ML models in HDD. Owing to the objective, the efficiency of 10 most widely used state-of-the-art ML models including LR, SVM, KNN, NB, DT, RF, GB, XGB and StGD have been evaluated and compared. To make an unbiased decision, each of these classical ML algorithms is evaluated with a different set of values of the parameters by doing parameter tuning. Table III listed the parameter values of each ML model with which each algorithm gives the best performance. Along with the parameter tuning, the concept of K-fold cross-validation is also used where different values of K are used. Among different values of K, for K = 10, almost all the ML models provide the highest classification performance.

E. *Evaluation Metrics*

For evaluating the performance of the ML models in HDD, we computed the Confusion matrix for every set of experiments. The confusion matrix comprises of True positive (TP), False positive (FP), True negative (TN) and False negative (FN). From the confusion matrix, the evaluation metrics like accuracy, sensitivity, specificity, precision, and F1-Score. The evaluation metrics like the AUC is also used for the evaluation of the performance of the ML models. The mathematical formula for each of these metrics are as follows:

$$Accuracy\ (Acc) = \frac{TP + TN}{TP + TN + FP + FN} X\ 100 \tag{1}$$

$$Sensitivity\ /Recall\ (Sens.) = \frac{TP}{TP + FN} X\ 100 \tag{2}$$

$$\text{Specifictity (Spec)} = \frac{TN}{TN + FP} X\ 100 \tag{3}$$

TABLE III
THE PARAMETER VALUES OF THE BEST FITTED ML MODELS

ML Algorithms	Parameter Details
LR	Penalty = 'L2', C = .01, Solver = liblinear
SVM	C = 1000, gamma = 1, kernel = rbf
KNN	metric='minkowski', n_neighbors = 15, weights='distance'
NB	Var_smoothing = 0.5336
ANN	Activation = 'relu', batch_size = 256, epoch = 30, Number of hidden layers = 3 with 45 neurons in 1st layer, 30 neurons in 2nd layer and 15 neurons in 3rd layer.
DT	Criterion = 'gini', max_depth = 12
RF	Criterion = 'entropy', max_depth = 12, n_estimator = 100
GB	Learning_rate = 0.1, max_depth = 7, n_estimator = 50
XGB	Max_depth = 10, min_child_weight = 2
StGD	Loss = Hinge, max_iter = 8, penalty = elastic net

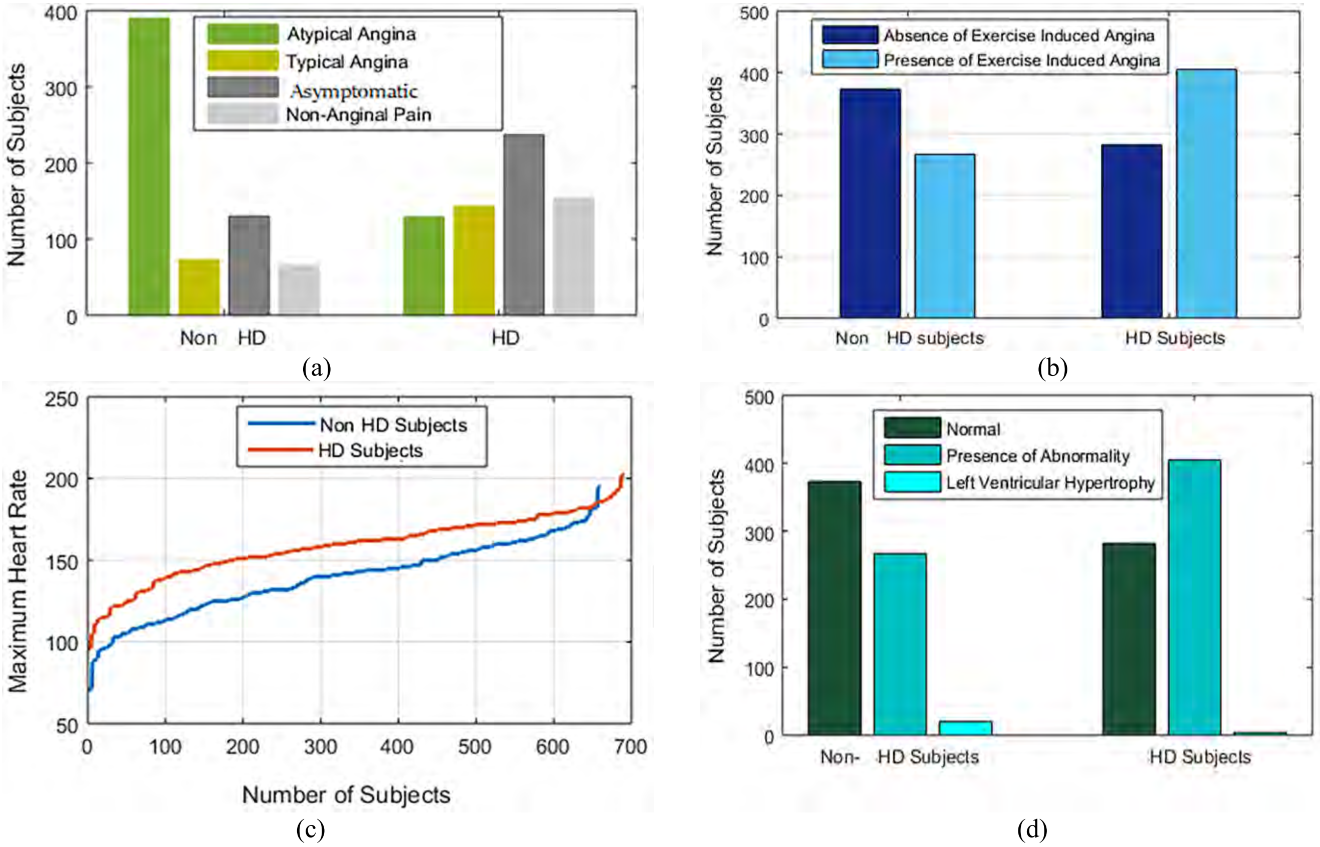

Fig. 2: (a) Number of subjects suffering from Various chest pain, (b) Difference in Exercise-induced chest pain in non-HD and HD group, (c) Difference in the range of Heart rate in non-DH and HD Group, and (d) Number of subjects with different status of ECG (taken in rest) in non HD and HD group. (b) A number of value differences between the non-HD and HD subjects.

$$\text{Precision (Prec)} = \frac{TP}{TP + FP}\ X\ 100 \tag{4}$$

$$\text{F1} - \text{Score} = \frac{2\ X\ Prec\ X\ Recall}{Prec + Recall} \tag{5}$$

Where, TP indicates the HD cases correctly classified as HD

TN indicates the non-HD cases correctly classified as non-HD

FP indicates the non-HD cases that wrongly classified as HD

FN indicates the HD cases that wrongly classified as non-HD.

F. *Designing of Hybrid ML System*

After computing the evaluation metrics from each set of ML models, the models are compared to select the models that provide classification accuracy above 90%. As illustrated in Fig. 3(a), out of ten ML models, only five models KNN, DT, RF, GB, and XGB provide both training as well as testing accuracy above 90%. Considering all these five models, an weighted voting based ensemble (WVE) classifier has been designed for HDD. To evaluate the performance of the proposed WVE system for HDD, the same abovementioned evaluation metrics have been used.

IV. Experimental Results and Discussion

The machine learning algorithms were executed using python programming language in windows operating system environment deployed in a computer system, Corei5 with 8GB Ram, 2.11GHz processor speed. All the necessary libraries are installed on python notebook to develop the proposed system.

A. *Results of ML models*

This section discusses the performances of various ML models in HD prediction. At first, we evaluated the performance of all ten ML techniques by using six metrics mentioned in the previous section. For better decision-making along with the classification accuracy on the test data, the classification accuracy on training data is also taken into account. As mentioned earlier, the performance of each classifier is obtained with a different set of parameter values. Out of these, the best parameters are considered for which each ML model gives the highest classification accuracy on the training dataset. The performance of each of the ML models with the best parameters (obtained with parameter tuning) are reported in Table IV. As illustrated in Table IV, out of all ML models, the highest training accuracy of 95.40% is obtained with KNN. Besides KNN, other models DT, RF, GB and XGB also provide better training accuracy of 94.56%, 94.23%, 95% and 95.20% respectively. Along with the train accuracies, the test accuracies of these models are also high. Among different models, the highest sensitivity is obtained with KNN and the highest specificity is obtained with RF model. Like sensitivity and specificity, the precision and F1-score of KNN, DT, RF, GB, and XGB are also very high compared to other state-of-the-art ML models LR, SVM, NB, ANN and StGD. The value of AUC is above 0.95 for all KNN, DT, RF, GB and XGB models. For highlighting the performance of these five models, the values of Train accuracy, test accuracy, sensitivity, specificity, precision, F1-Score and AUC are marked in Boldface font in Table IV.

B. Results of Ensemble Classifier

After identifying the best performing models in HD prediction, next phase is to design a hybrid system by combining the decision of each of these ML models. Unlike a single model, a hybrid ML model will improve the HDD system performance. The basic idea is to combine the decisions of multiple ML models by using the voting policy. However, Voting Ensemble (VE) gives equal importance to each of the contributing model for which sometimes the system performance does not get improved. Considering this, in this work we are using WVE. As reported in Table IV, each of the classifiers are not equally capable in HD classification. Some models are better than other. So, based on this key idea, different weight coefficients are assigned to each of the classifiers. The values of these weight coefficients are in between 0 and 1. However, selection of appropriate and unbiased weight is a challenging task. In this work, for assigning weights to five best performing classifiers, we have used the concept of validation set. The whole experimental dataset is divided into training and testing set in a ratio of 7:3. Then this training set is further split into training and validation set in a ratio of 6:4. Then, the classification accuracy of each of KNN, DT, RF, GB and XGB is measured on the validation dataset. These accuracies are used as weight coefficient for each of these classifiers. The set of weights for KNN is 0.8956, for DT is 0.8874, for RF is 0.8984, for GB is 0.8462 and for XGB is 0.8709. With these weight values, the ensemble classifier provides an accuracy of 96.15% on train dataset and 95.64% on test dataset. Like the improved accuracy, the values of other parameters (except specificity) are also better for the proposed WVEHDD system as illustrated in Fig. 4. The values of the sensitivity, specificity, precision, F1-score and AUC are 0.9896, 0.924, 0.96, 0.96 and 0.99 respectively. As illustrated in Fig. 4, in contrary to the other conventional ML models, the proposed hybrid models give better training and testing accuracy. Like training and testing accuracies, the values of other parameters are also get improved in our proposed WVEHDD system.

C. Comparison of Proposed WVEHDD with other existing methods

To evaluate the potential of the proposed WVE based HD detection system, we need to compare our work with some other relevant works. As reported in literature survey section, a lot of works were already done for HDD, but in almost all the methods, the experimental dataset was very small containing only 303 instances. Very few works were done that considered a large experimental dataset containing at least more than 1000 samples. Moreover, as reported in literature almost all works considered all the features of the dataset, i.e. all 13 features. Considering these factors, we have designed an WVE based HDD system that considers only 4 statistically significant features for HD prediction and the efficiency of the proposed method was evaluated on an experimental dataset containing 1296 samples. A comparative overview of other works and ours is provided below.

Bashir et al. [9], 2016 reported the highest classification accuracy of 81.82% from an ensemble classifier. Like Bashir et al. [9], S. Kumar et al. [7] also reported the highest classification accuracy of 86.8%. In [15], the highest accuracy of 90% was obtained by using an ensemble technique. K. Polat et al. [10] reported a classification accuracy of 96.37% on a small experimental dataset comprising of 267 samples. Moreover, like the others, they also considered all 22 features of the dataset. S. A. E. Mienye et al. [11] reported a classification accuracy of 91% using 16 features of their experimental dataset comprising of 4238 samples. In [12], Dan Gan et al. reported a classification accuracy of 80.27%. D. Shah et al. obtained a classification accuracy of 91.78% using the KNN classifier. D. K Plati et al. [16] also reported the highest classification accuracy of 91.23% on an experimental dataset containing 487 instances. In [14], Muhammad Waqar et al. reported the highest classification accuracy of 96.1% in an experimental dataset of 303 samples.

TABLE IV
PREDICTION PERFORMANCES OF THE PROPOSED WVEHDD AND OTHER EXISTING ML BASED HD DETECTION

ML Models	Train Accuracy	Test Accuracy	Sensitivity/ Recall	Specificity	Precision	F1-Score	AUC
LR	75.10%	78.10%	0.73	0.82	0.78	0.78	0.83
SVM	80.3%	79.20%	0.67	0.87	0.79	0.79	0.87
KNN	**95.40%**	**93.10%**	**0.95**	**0.91**	**0.93**	**0.93**	**0.98**
NB	75.90%	79.60%	0.73	0.85	0.80	0.79	0.84
ANN	74.70%	80.10%	0.77	0.79	0.78	0.78	0.84
DT	**94.56%**	**90.40%**	**0.93**	**0.89**	**0.93**	**0.91**	**0.98**
RF	**94.23%**	**93.10%**	**0.94**	**0.93**	**0.93**	**0.93**	**0.99**
GB	**95.00%**	**92.70%**	**0.93**	**0.92**	**0.93**	**0.93**	**0.97**
XGB	**95.20%**	**92.70%**	**0.93**	**0.92**	**0.93**	**0.93**	**0.97**
StGD	72.56%	71.26%	0.70	0.74	0.76	0.76	0.75

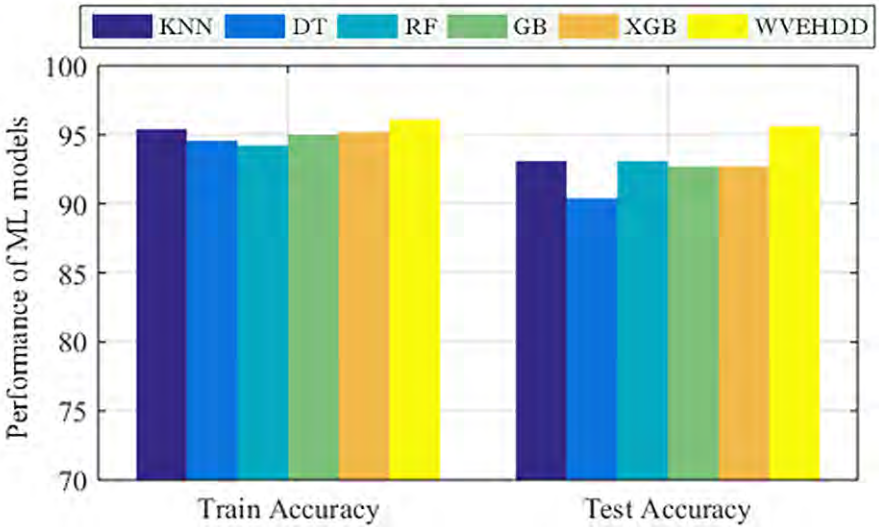

Fig. 3. Comparison of the Training & Testing accuracy of five best fitted models with the proposed WVEHDD system.

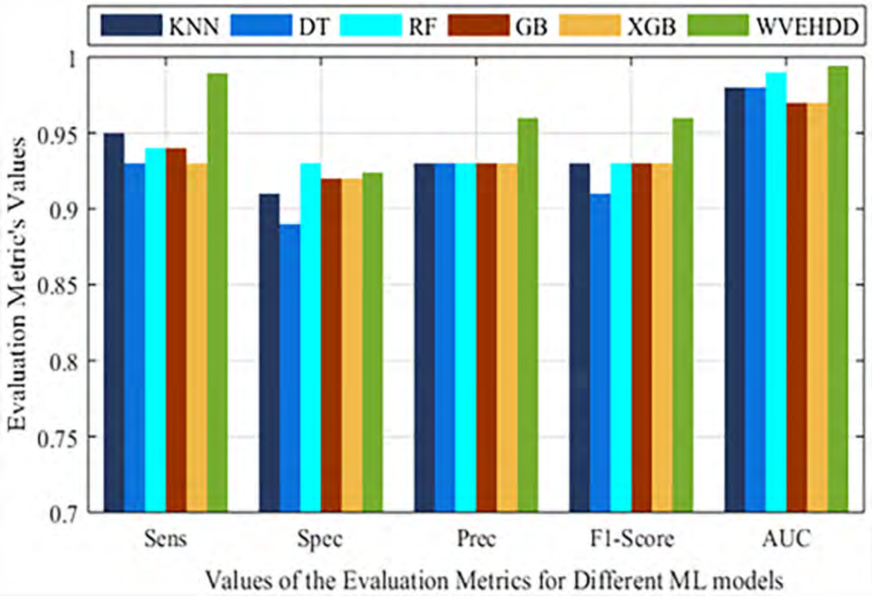

Fig. 4. Comparison of the Training & Testing sensitivity, specificity, precision, F1-Score, AUC of five best fitted models with the proposed WVEHDD system.

In comparison to these methods, our proposed WVEHDD method reported a train accuracy of 96.23% and test accuracy of 95.24% in an experimental dataset comprising of much more samples than the other existing methods. Moreover, unlike the other methods, our proposed WVEHDD method considers only four statistically significant feature to represent the HD. Considering all these parameters, we can conclude that our proposed WVEHDD system outperforms the other existing methods. The summary of each of the above works are provided in Table V.

V. Conclusion

Because of the availability of a significant amount of publicly available medical data, the researchers are using various data-mining, data analysis approaches for disease prediction and categorization. One of the biggest causes of death in the world is HD, and an extensive amount of research is being carried out to develop intelligent disease diagnostics. Although several research works are there in literature, almost all the works were based on small experimental dataset and all the features of the datasets were used. Hence, the paper emphasizes on considering a larger experimental dataset with fewer features to design an intelligent heart disease diagnostic system having the same or better potential as in the earlier works. The paper proposed a weighted vote-based ensemble classifier for HDD. In order to design the WVE classifier, the classification performance of ten ML algorithms: LR, SVM, KNN, NB, DT, RF, GB, XGB and StGD are evaluated by using six most widely used evaluation metrics: accuracy, sensitivity, specificity, precision, F1-score, and AUC. Based on the values of these evaluation metrics, five ML techniques:

TABLE V
PREDICTION PERFORMANCES OF THE PROPOSED WVEHDD AND OTHER EXISTING ML BASED HD DETECTION

Authors	**Dataset**	**Fs no.**	**Methods**	**Performance Measures**
Bashir et al. [9], 2016	CHD (303)	13	Ensemble approach (NB, DT (Gini Index), DT (Information Gain), instance-based-learner, SVM)	Acc - 81.82%, Sensitivity –73.68%, Specificity –92.86%, F-measure - 82.17%
S. Kumar et al. [7], 2018	CHD (303)	13	DT (C4.5) & GA	Acc - 86.8%
R. Atallah et al. [15], 2019	CHD (303)	13	SGD, KNN, RF, LR	SGD - 88%, KNN - 87% RF - 87%, LR - 87%, Ensemble - 90%
K Polat et al. [10], 2018	SPECT dataset (267)	22	LDA, KNN, SVM, and RF	Acc - 96.637%, Precision –97.10%, Recall –96.60%, AUC – 99.00% F-measure -96.70%
S. A. E. Mienye et al. [11], 2020	Framingham Heart Dataset (4238)	16	Improved sparse autoencoder (SAE) + Softmax	Acc - 91%, Precision –93%, Recall –90%, F-measure -92%
Dan Gan et al. [12], 2020	CHD (303)	13	AdaCTANBN	Acc – 80.27%, Precision –88.873%,
D. Shah et al. [13], 2020	CHD (303)	13	NB, DT, KNN and RF	Train Acc – 91.78% (KNN) Test Acc – 88.16% (NB)
D. K Plati et al. [16], 2021	Private dataset: 487 instances	--	RF, Rotational Forest, NB, KNN, SVM, LMT, BN	Acc – 91.23%, Sens – 93.83%, Spec – 89.62
Muhammad Waqar et al. [14], 2021	CHD (303)	13	KNN, SVM, LR, RF, NB, ANN, Ensemble classifier	Acc – 96.1%, Prec – 95.7% Rec – 95.7%, F1 Score – 96%
Proposed WVEHDD	CHD + HDD (1296)	4	LR, SVM, KNN, NB, DT, RF, GB, XGB and StGB, Ensemble Classifier (KNN, DT, RF, GB, XGB)	**Train Accuracy – 96.23%,** **Test Accuracy – 95.24%,** Sens. – 98.30%,Spec – 92.6% Prec – 95% , F1-Score – 95% , AUC – 0.99.

KNN, DT, RF, GB, and XGB are found to be much better than the other five classifiers. Then the weighted votes of these five individual classifiers are combined to get the final decision. The performance of the ensemble classifier has been evaluated using the six evaluation metrics. The values of each of these six parameters shows the high potential of the proposed WVEHDD system. The comparative study of the proposed WVEHDD method with the other relevant works also reveals that the proposed WVEHDD system outperforms the existing systems.

In the future study, it is aimed to generate some derived features from the raw features to improve the accuracy of the proposed system by using the same ensemble classifier.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ACKNOWLEDGMENT

The work presented herein was conducted at the School of Science and Technology of The Neotia University, Diamond Harbour Rd, Sarisha, West Bengal 743368, West Bengal, India.

REFERENCES

[1]. Cardiovascular disease,
Available at: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[2]. Ahsan, M. M., & Siddique, Z. Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, pp. 102289, 2022; https://doi.org/10.1016/j.artmed.2022.102289

[3]. Murray CJ, Lopez AD. Alternative projections of mortality and disability by cause 1990–2020: Global Burden of Disease Study. *The Lancet*, vol. 349, pg. 1498–1504, 1997; https://doi.org/10.1016/S0140-6736(96)07492-2.

[4]. Heart Disease Cleveland UCI,
Available at: https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci

[5]. The Essential Guide to Quality Training Data for Machine dLearning,
Available at https://www.cloudfactory.com/training-data-guide#:~:text=The%20quality%20and%20quantity%20of%20your%20training%20data,a%20model%20trained%20on%20data%20from%2010%2C000%20transactions

[6]. An Introductory Guide to Quality Training Data for Machine Learning,
Available at: https://www.v7labs.com/blog/quality-training-data-for-machine-learning-guide

[7]. Kumar S, Sahoo G. Enhanced decision tree algorithm using genetic algorithm for heart disease prediction. *International Journal of Bioinformatics Research and Applications*, vol. 14(1/2), pp. 49-69, 2018, 10.1504/IJBRA.2018.10009164.

[8]. Heart Disease Dataset:
Available at: https://archive.ics.uci.edu/dataset/45/heart+disease

[9]. Bashir S, Qamar U, & Khan F H. A multicriteria weighted vote ‐ based classifier ensemble for heart disease prediction. *Computational Intelligence*, vol. 32(4), pp. 615-645, 2016; https://doi.org/10.1111/coin.12070.

[10]. Polat K. Similarity-based attribute weighting methods via clustering algorithms in the classification of imbalanced medical datasets. *Neural Computing and Applications*, vol. 30, pp. 987-1013, 2018; https://doi.org/10.1007/s00521-018-3471-8

[11]. Ebiaredoh-Mienye SA, Esenogho E, & Swart TG. Integrating enhanced sparse autoencoder-based artificial neural network technique and softmax regression for medical diagnosis. *Electronics*, vol. 9(11), pp. 1963, 2020; https://doi.org/10.3390/electronics9111963

[12]. Gan D, Shen J, An B, Xu M, & Liu N. Integrating TANBN with cost sensitive classification algorithm for imbalanced data in medical

diagnosis. *Computers & Industrial Engineering*, vol.140, pp. 106266, 2020, https://doi.org/10.1016/j.cie.2019.106266.

[13]. Shah D, Patel S, & Bharti SK. Heart disease prediction using machine learning techniques. *SN Computer Science*, vol. 1, pp. 1-6, 2020, https://doi.org/10.1007/s42979-020-00365-y.

[14]. Waqar M., Dawood H, Dawood H, Majeed N, Banjar A, & Alharbey R. An efficient SMOTE-based deep learning model for heart attack prediction. *Scientific Programming*, vol. 2021, pp. 1-12, 2021; https://doi.org/10.1155/2021/6621622

[15]. Atallah R, & Al-Mousa A. Heart disease detection using machine learning majority voting ensemble method. In 2019 2nd international conference on new trends in computing sciences (ictcs), pp. 1-6, IEEE, 2021, 10.1109/ICTCS.2019.8923053.

[16]. Plati D K, Tripoliti E E, Bechlioulis A, Rammos A, Dimou I, Lakkas L, et al. A machine learning approach for chronic heart failure diagnosis. *Diagnostics*, vol. 11(10), pp. 1863, 2021; https://doi.org/10.3390/diagnostics11101863.

[17]. The Wilcoxon Rank Sum Test. Available at: https://data.library.virginia.edu/the- wilcoxon-rank-sum-test/

Usha Rani Gogoi is working as Assistant Professor in the Computer Science & Engineering department of The Neotia University. She received her Bachelor of Engineering (B.E.) degree with honor from Guwahati University in 2011. She received her Master of Technology (M.Tech.) degree in Computer Science and Engineering from Tripura University (A Central University) in 2014 and held the First class first position. She received her Ph.D. degree in Computer Science & Engineering from Tripura University (A Central University), Tripura, India in 2020. Dr. Gogoi pursued her Ph.D. degree under the DST Inspire Fellowship scheme. She has around 13 research publications including Science Citation Indexed Journals. Her areas of research interest are medical image processing, artificial intelligence, machine learning, pattern recognition, computer-aided system design, and soft-biometric.

i