Heart Disease Prediction Model based Ongradient Boosting Tree (GBT) Classification Algorithm

R. Bhuvaneeswari, P. Sudhakar, G. Prabakaran

Abstract--- Recently, Heart disease (HD) is the main cause of increasing death rate all over the world. Data classification is a crucial task in the medical field which assists the physicians to predict the diseases. Recently, machine learning (ML) algorithms have been employed to classify the data in the medical field. The data complexity and quantity needs to be examined and managed to transform the efficient and accurate HD diagnosis. In this paper, a gradient boosting tree (GBT) based classifier or gradient boosting classifier (GBC) model to predict the HD efficiently. Besides, a set of extensive experiments were carried out using Staglog and Cleveland heart disease dataset. The experimental values ensured the superiority of the GBT classifier based on several performance measures.

Keywords--- Heart Disease, Machine Learning, Classification, Gradient Boosting Tree.

I. INTRODUCTION

Heart disease (HD) is a commonly occurring disease in both men and women. It inhibits the heart in satisfying the bodies circulatory demands as it affects the ventricle ability to eject or fill blood. Fatigue, ankle swelling, and breathlessness are the symptoms that might be followed by signs; for instance, pulmonary crackles, peripheral edema, and elevated pressure in jugular venous which are caused through functional and structural cardiac or non-cardiac abnormalities. It is a severe condition linked to high mortality and morbidity rates. Also, based on the European Society of Cardiology (ESC), 26 million adults are diagnosed globally with HD whereas 3.6 million are diagnosed newly every year. In the first year, patients at 17–45% suffer from HD dies whereas the rest dies in 5years. Around 1–2% out of all the expenditure in healthcare falls under HD management, with many associated and recurring hospital admissions [1–3]. The enhanced healthcare costs, condensed quality of life (QoL), increased occurrence, frequent hospitalizations, early mortality have converted HD to a severe epidemic worldwide and drawn attention to the requirement of earlier diagnosis and efficient treatment.

Medical diagnosis in clinical practice involves careful physical and history examinations. It is aided through ancillary tests, like chest radiography, echocardiography [4], blood tests and electrocardiography. It has created data combination through the above diagnosis results in numerous condition formulations that determines the existence of HD [5]. Using the guidelines of classification systems, they classify the HD severity by employing either American College of Cardiology/American Heart Association (ACC/AHA) Guidelines or New York Heart Association (NYHA), as this enables to decide the highly suitable treatment that has to be followed[6]. The data complexity and quantity needs to be examined as well as managed to transform the efficient and accurate HD diagnosis for accessing the therapeutic regimens to complicated and challenging tasks. The reason under the high improvement of machine learning (ML) methods application to predict, classify and analyze medical data is to diagnose the HD as early as possible thereby reducing the patient conditions worsening, enhances QoL and to minimize the medical expenses. Over the methods of data mining, the classification techniques draw the attention of researchers. Precise disease classification allows identifying the disease subtypes, stage or etiologywhich enables the suggestions and treatments in assessing the patient's progress.

Various methods of data mining were aimed at HD and have been used to distinguish the HD patients to identify various HD sub-kinds and to compute HD severity. At a later stage, if the heart failure is detected, the methods of data mining can be useful wherever the therapeutic advantages for aid and survival prospect are constrained as it enables for timely prediction of morbidity, the risk of readmission and mortality. Subjects' health record, Clinical history data, physical examination results, electrocardiogram (ECG) analysis results,subjects' health record, laboratory data, presenting symptoms, expressing demographic information are the data will be recorded. Extensive studies of the problem that are addressed above by the implication of machine learning methods are projected in this work. The method to detect heart failure [1] is a non-acute setting in order to ESC guidelines. Primarily, the HD probability depends on previous patient clinical records, physical examination, symptoms present and remaining ECG is computed. HD is improbable if the entire elements are usual. The Natriuretic plasma Peptides has to be measured if one element is weird. This metric enables the expert to recognize the patients who are in requirement of echocardiography.HD detection is a two class classification issue wherever the classifier output is present or not.

More studies aim at heart rate variability (HRV) utilization; a metric used to divide a subject as a patient or standard with HD.

Manuscript received August 19, 2019.

R. Bhuvaneeswari, Research Scholar, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Annamalai University. (e-mail: anand.anandrajendran@gmail.com)

P. Sudhakar, Assistant Professor, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Annamalai University. (e-mail: kar.sudha@gmail.com)

G. Prabakaran, Assistant Professor, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Annamalai University. (e-mail: gpaucse@yahoo.com)

The significantvariations among the techniques are relative to a feature of HRF that are used to detect HD.A scoring model is projected through Yang et al. 2010 [7], used to analyze the seriousness of the disease and HD detection. Two models of Support Vector Machines (SVM) were developed. The primary model finds the absence or presence of HD. The next model divides the patients to HD-prone or Healthy group. The SVM model output is mapped toward a scoring rate. When the scoring rate generated through the first model output by mapping is lower than 4, then the subject depends on the non-HD group. To detecting HD, neural networks (NN) have been used [8] with a 40 subject set. For every subject, age, smoking habit, annotation as patient or normal, gender, and blood pressure are available. Among the total of 40 subjects, 38 were classified correctly that results to False Positive Rate of 9.00%, Recall of 95.00%, Area Under Curve (AUC) of 95%, True Positive Rate 95.00%, Precision 95.00% and F-measure 94.00%.

To distinguish the patient suffering from dyspnea and congestive heart failure (CHD), Son et al. 2012 [4] reviewed the study of 72 variables discrimination power. For feature space reduction, logistic regression (LR) and rough sets were used. A classification based on DT was employed. The simulation results demonstrate the sensitivity 97.2%, positive predictive value 97.2%, area under ROC curve 97.5%, accuracy 97.5%, specificity 97.7% andnegative predictive value 97.7%. For long-term ECG time sequence, Random Forests (RF) is used by Masetic et al. 2016 [9] for the detection of HD. From Beth Israel Deaconess Medical Center (BIDMC), ECG signals were accessed and PTB Diagnostic ECG, as well as Congestive Heart Failure databases, were available at PhysioNet [10] when the actual heartbeats were derived from MIT–BIH Arrhythmia database comprising 13 subjects. By employing the method of autoregressive Burg, one feature is derived from ECG. Over the similar dataset, the methods like SVM, k-Nearest Neighbors (k-NN), C4.5 and Artificial Neural Networks (ANN) were examined using specificity, F-measure, sensitivity, accuracy and ROC curve. Because of the efficient accuracy, RF has been selected in the classification of the subject as CHD or normal.

Another study is made to try the prediction of HD presence. Beforethe actual date of clinical diagnosis, Wu et al. 2010 [11] designed anHD detection for higher than six months. The electronic health data records involved health behavior, clinical diagnosis, laboratory data, demographic, use of care, clinical measures, and prescription orders for anti-hypertensive data. The data was demonstrated through 179 independent variables. For earlier HD prediction, the researchers compared Boosting, LR models and SVM. Feature selection wasmade, before classifiers application. Based on the classifier, the various selection process was used. Selection of variable depends on reducing the Bayesian information criterion (BIC) and Akaike information criterion (AIC), in case of SVM, L1-norm variable selection method was employed. AUC has been measured, and the outcomes denote that AUCs were same as boosting and LR.

A multi-level risk examination of constructing HD was projected through Aljaaf et al. 2015 [2]. There are five risk levels present in which the projected model can predict by employing C4.5 DT classifier.The dataset used is the Cleveland Clinic Foundation for heart disease. With three new attributes namely physical activity, smoking and obesity, the researcher improved the dataset. A total of 35 samples were collected for physical activity, 54 samples were gathered for smoking, and 160 were collected for obesity. A process of 10-fold cross-validation is performed for the analysis using C4.5 classifier. The precision on the whole for the projected method is 86.30%. For chronic HD diagnosis, a computer assisted system has been projected [12]. It is tested and trained using heart sound and cardiac reverse features with Least Squares SVM (LS-SVM). The LS-SVM classifier results were examined with the methods like Hidden Markov Models and ANN denoting the LS-SVM approach superiority.

Though different classifier models have been presented, there is still a need to enhance the performance. In this paper, a gradient boosting tree (GBT) based classifier model on predicting the HD efficiently. Also, a set of extensive experiments were carried out using Staglog and Cleveland heart disease dataset. The experimental values ensured the superiority of the GBT classifier based on several performance measures.The remaining part of the paper is arranged as follows. Section 2 briefs the presented GBC for HD prediction. Section 3 validates the results and Section 4 concludes the study.

II. PROPOSED GBT CLASSIFIER FOR HD PREDICTION

The overall process of the GBT classifier for HD prediction is given in Fig. 1. Initially, the dataset is read and is pre-processed for futhrer operations. The data preprocessing involves two sub-processes namely format conversion and missing values replacement. Once the format is converted and the missing values are replaced, the classification process will be initiated. Then, the data calssificaiton takes palce by GBT classifiers and then the performance is evaluated under several evaluationparamaters.

2.1 The Tobit model

Auxiliary data for the favored class can be integrated with the binary data utilizing censored regression approaches. The two-sided versions of the Tobit model [13] is a commonly employed censored regression model. The Tobit model postulates that there survive a latent variable Y^* which follows, conditional on some covariates $X = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$, a Gaussian distribution:

$$Y^*|X \sim N(F(X), \sigma^2) \quad (1)$$

Mean $F(X)$ is considered as a depend linearly on the covariates X by $F(X) = X^T\beta$, where $\beta \in \mathbb{R}^p$ is a collection of coefficients.

This latent variable $Y^*$is noted that when it lie in the interval $[y_l, y_u]$. Else, one notices y_l or y_u based on the fact that latent variable is the lesser the lower threshold $y_l$or above the upper threshold y_u, correspondingly.

It is denoted that Y is the observed variable and can be expressed as

$$Y = \min(\max(y^*, y_l), y_u) \quad (2)$$

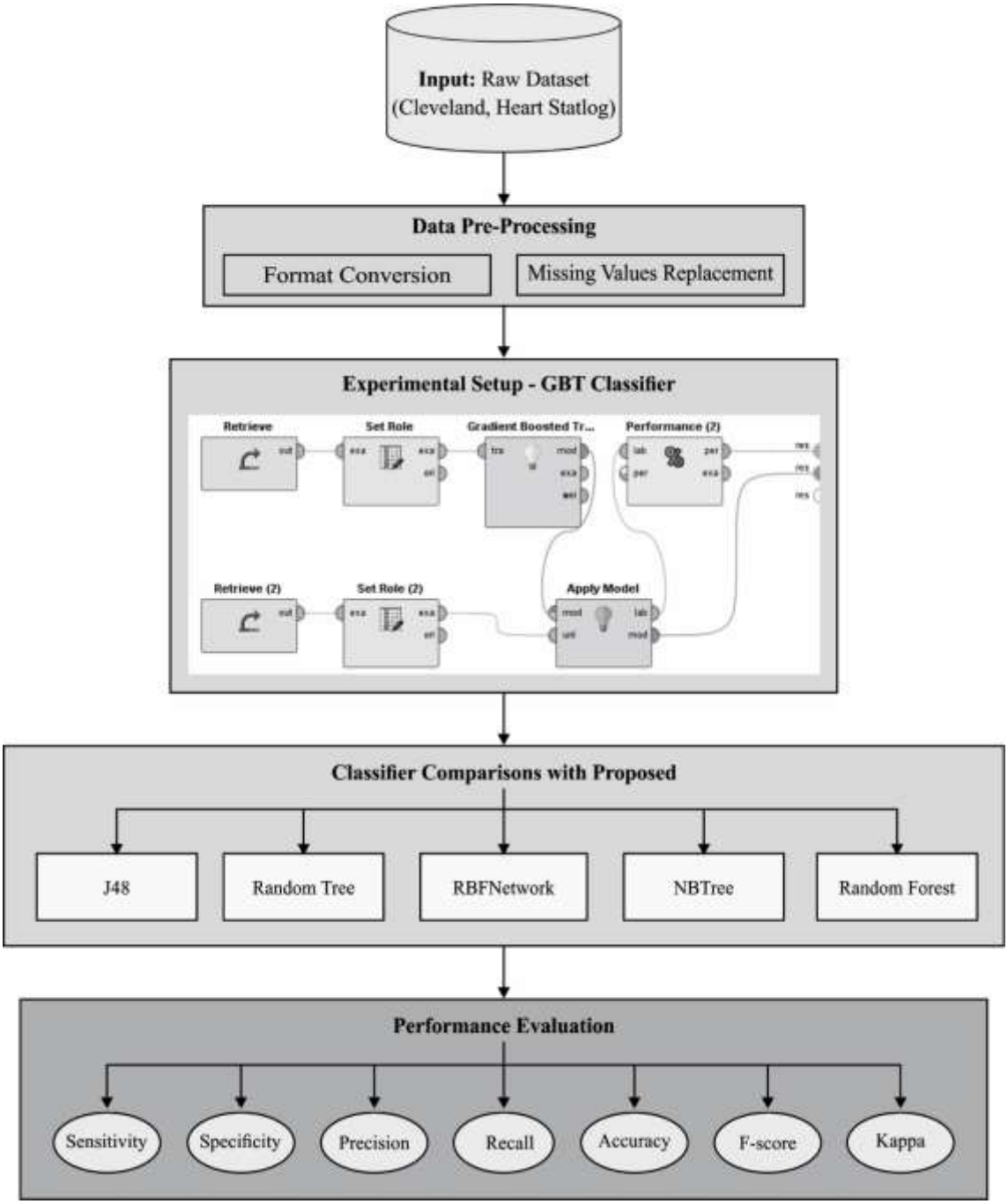

Fig. 1: Overall process of the proposed method

It is noted that the one-sided Tobit model is attained in particular cases by letting one of the boundaries $y_l$or y_u converges to minus or plus infinity. Though the Tobit model is represented as a censoring strategy and it can be employed to real censored data, however, in diverse situations, the data comprises of continuous as well as the discrete point at the borders. This includes fractional response data, loss given default, rainfall, or default prediction as in this article. The latent variable Y^* can often be interpreted as a potential which indicates how likely the event under consideration is to occur. In this case, Y^* is interpreted as a default potential, and a default appears when the potential Y^* crosses a particular threshold. Default events represents the case $Y^* \geq y_u$, and the observed data is identified with $Y = y_u$. The non-default cases correspond to $Y^* < y_u$, and the secondary variable is identified with the observed variable $Y = Y^*$ in this case.

The concentration is generally not on econometric or statistical approaches instead of the minimization of loss functions. From this point of view, the Tobit model obtains a proper symmetric loss function. In addition, it is not necessary to consider that the observed default data and the secondary data are created by Tobit model. It is also considered that the Tobit likelihood as a tool which enables to combine the binary default data with the secondary data.

2.2 Boosting

A relatively uncertain statement of the Tobit model is the linear function that integrates a collection of covariates to a linear predictor. This statement is relaxed here by the application of gradient tree boosting to the Tobit model, and the resultant one is called as 'Grabit' model. Boosting utilizes high familiarity in a different region because of higher predictive accuracy on a broad range of dataset, e.g., [14]. It integrates many relatively simple approaches which are also known as learners comprising of regression trees. Boosting is initially employed in ML for classifying data. The main contribution is a statistical perception of boosting as stage-wise optimization of a risk functional.

The concept of boosting method is developed by [15]. A response variable Y and a vector of covariates $X \in \mathbb{R}^p$ and observe data $(y_i, x_i), i = 1, \ldots, n$. The aim of boosting is to discover a minimization function $F*(\cdot)$ of the empirical loss $R^e(F)$

$$F^*(\cdot) = \underset{F(\cdot) \in \Omega_s}{argmin} R^e(F) \quad (3)$$

$$= \underset{F(\cdot) \in \Omega_s}{argim} \sum_{i=1}^{n} L\big(y_i, F(x_i)\big) \quad (4)$$

where $F(\cdot) : \mathbb{R}^p \to \mathbb{R}$ are functions which is mapped to Y and L is a properly selected loss function like squared loss $L(y,F) = (y - F)^2/2$or the negative Tobit log likelihood. In boosting, single restrict the functions $F(\cdot)$ to lie in the span $\Omega_S = span(S)$ of a set S of so-called base learners $h(x; a^{[m]})$. Specifically, boosting construct an ensemble

$$F(x) = F^{[0]} + \sum_{m=1}^{M} \rho^{[m]} h\big(x; a^{[m]}\big) \quad (5)$$

where it is considered that $h(x; a^{[m]})$ are regression trees with parameters $a^{[m]}$, $F^{[0]}$ is a constant, $\rho^{[m]} \in \mathbb{R}$, and M indicates the number of boosting iterations or trees. The boosting method repeatedly identifies $F^*(\cdot)$ by the use of functional gradient descent method. Indicating the present estimate for $F^*(\cdot)$ by $F^{[m-1]}(\cdot)$, an update from$F^{[m-1]}(\cdot)$ to $F^{[m]}(\cdot)$, is attained by the initial calculation of the negative gradient $-\frac{\partial L(y_i, F(x_i))}{\partial F}\Big|_{F=F^{[m-1]}}$, and then resembling this gradient with a base learner $h\big(x; a^{[m]}\big)$. When the trees are employed as base learners and the second derivative of the loss function $L(y,F)$ is a non-constant and non-zero value. It recommends to perform an extra step of Newton's approach to identify the leaf values. When the hybrid gradient-Newton approach is presented where the tree partition are learned by the use of gradient descent and the leaf values are learned by the use of Newton's method. Furthermore, a shrinkage factor $\nu, 0 < \nu \leq 1$ is commonly employed to update step:

$$F^{[m]}(x) = F^{[m-1]}(X) + \nu\rho^{[m]} h\big(x; a^{[m]}\big) \quad (6)$$

This parameter ν performs as a regularization parameter. It is also noted that the usage of a shrinkage factor reduces the overfitting and leads to maximum predictive results.

2.3 The Grabit model

Grabit model is employed by improving the Tobit model by the use of GB with trees as base learners. Though nonlinearities and interactions are defined in different ways, boosting with trees offers flexibility and depends on some considerations. Notably, it depicts better classifier results on a diverse dataset. Rather than considering a linear function for the mean function $F(X)$ of the latent variable Y^* , the Grabit approach make use of flexible function $F(\cdot) : \mathbb{R}^p \to \mathbb{R}$ that comprises of an ensemble of regression trees. An approximation of this function is attained by employing boosting with regression trees as base learners. Particularly, negative log-likelihood of the Tobit model is applied as the loss function $L(y,F)$:

$$\mathrm{L(y, F)} = -\log\big(\mathrm{f_{F,\sigma}(y)}\big) \quad (7)$$

The boosting method operates on repeatedly fitting a regression tree $h(x_i, a^{[m]})$ as a least squares approximation to the so-called pseudo responses $\tilde{y}_i$ as given below

$$\tilde{y}_i = -\frac{\partial L(y_i, F)}{\partial F}\Big|_{F=F^{[m-1]}(x_i)} \quad (8)$$

Then, the optimum $\rho^{[m]}$ is attained by the minimization of the empirical loss as represented below.

$$\rho^{[m]} = \underset{\rho}{argmin} \sum_{i=1}^{n} L\Big(y_i, F^{[m-1]}(x_i) + \rho h\big(x_i, a^{[m]}\big)\Big) \quad (9)$$

A second order Taylor approximation for $\sum_{x_i \in R_j^{[m]}} L\big(y_i, F^{[m-1]}(x_i) + \gamma\big)$ is employed $F^{[m-1]}(x)$, and identify the $\gamma_j^{[m]}$so that the approximation is minimized. It represents a single Newton-Raphson step as defined below.

$$= -\sum_{x_i \in R_j^{[m]}} \frac{\partial L(y_i, F)}{\partial F}\Big|_{F=F^{[m-1](x_i)}} \Big/ \sum_{x_i \in R_j^{[m]}} \frac{\partial^2 L(y_i, F)}{\partial^2 F}\Big|_{F=F^{[m-1]}(x_i)} \quad (10)$$

$\gamma_j^{[m]}$

The gradient descent step can be used to identify the tree structure, i.e., the partition of the space, and a Newton update step for learning the leaf values. The parameter σ is considered as a known parameter. The parameter σ is selected using cross-validation or determined by the use of profile-likelihood.

2.4 Choice of tuning parameters

The Grabit algorithm has numerous tuning parameters tree count M, the shrinkage factor ν, and tree depth T. Moreover, the standard deviation σ of the latent variableY^*. The shrinkage factor ν and the tree M control the amount of regularization. Past research has shown that the predictive accuracy of boosting algorithms is generally superior when choosing smaller values for ν. The depth of the trees T controls the degree of interaction among the covariates X. A tree of depth T can maximally have interactions of order $T - 1$. The parameters ν, M, and T is selected using cross-validation or information criterion. The parameter σ is selected by increasing the profile likelihood or cross-validation. The profile log-likelihood function for σ is represented by the negative empirical loss as given by

$$\ell(\sigma) = -R^e\big(\hat{F}_\sigma, \sigma\big) \quad (11)$$

$$\hat{\sigma} = \underset{\sigma}{argmax}\, \ell(\sigma) \quad (12)$$

For avoiding the issues by the use of negative values, one can reparametrize σ by $\phi = log(\sigma) \in \mathbb{R}$, identify $\hat{\phi} = argmax_\phi \ell\big(e^\phi\big)$and set$\hat{\sigma} = e^{\hat{\phi}}$.

A computationally quicker but significantly less precise alternative is to do a grid search over a grid $\{\phi_1, \ldots, \phi\}_k$, where K is the number of grid points.

III. EXPERIMENTAL RESULTS

To examine the proposed GBC for the HD prediction, numerous analyseswere performed, and the results of the datasets are provided below. GBC is used to classify the samples of the datasets. Dataset description, frequency distribution, class distribution, confusion matrix and performance evaluation with various other classifier are given in the following sections.

3.1 Dataset Description

For the performance validation of GBC, two benchmark datasets namely Heart-Statlog [16] and Cleveland[17] were used. Heart-Statlog containstwo classes, 270 instances, 13 number of features, 55.50% of instances falls under the absent samples, the percentage of 44.50% falls under present samples. For Cleveland dataset, it is composed of 2 classes, 303 instances, 13 number of features, the percentage of 55.50% falls under the absent samples, the percentage of 44.50% falls under present samples. The dataset descriptions are given in Table 1. The attribute description is given in Table 2 which it comprises of dataset descriptions, data type, and attribute name. The data type is nominal and real for most of the attributes. There exist two classes to predict the presence and absence of HD. For example, a few names of the attributes are restbps, age, cp, sex, and so on. For all attributes, Figs. 2 and 3 demonstrate the frequency distribution of heart disease datasets. Fig. 4 gives the class distribution of heart disease datasets.

Table 1: Dataset Description

Description	Heart-Statlog	Cleveland
Number of Instances	270	303
Number of Features	13	13
Number of Class	2	2
Percentage of Present Samples	44.50%	54.13%
Percentage of Absent Samples	55.50%	45.87%
Data sources	[16]	[17]

Table 2: Attributes Description

Attribute	Data Type	Attribute description
age	Real	Age (in years)
sex	Binary	0—female, 1—male
cp	Nominal	Chest pain type (1—typical angina, 2—atypical angina, 3—nonangina, 4—asymptomatic)
restbps	Real	Resting blood pressure (mm of Hg)
chol	Real	Serum cholesterol (mg/dL)
fbs	Binary	Fasting blood sugar >120 mg/dL (1—true, 0—false)
restecg	Nominal	Resting electrocardiographic results (0—normal, 1—having ST-T wave normality, 2—probable/defined left ventricular hypertrophy)
thalach	Real	Maximum recorded heart rate
exang	Binary	Angina induced by exercise (1—yes, 0—false)
oldpeak	Real	ST depression tempted by workout comparative to rest
slope	Nominal	Slant of the peak exercise ST segment (1—upsloping, 2—flat, 3—downsloping)
ca	Real	Major vessels colored by fluoroscopy
thal	Nominal	3—normal, 6—fixed defect, 7—reversible defect
class	Binary	Represent present or absence of heart disease (1—absence, 2—presence)

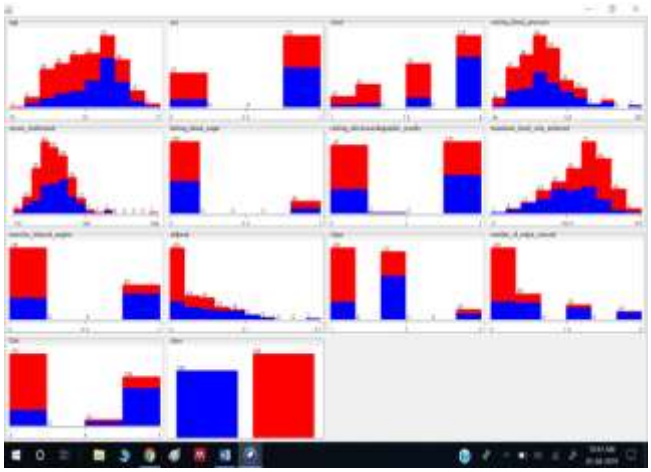

Fig. 2: Frequency Distribution of Heart Statlog Dataset for all Attributes

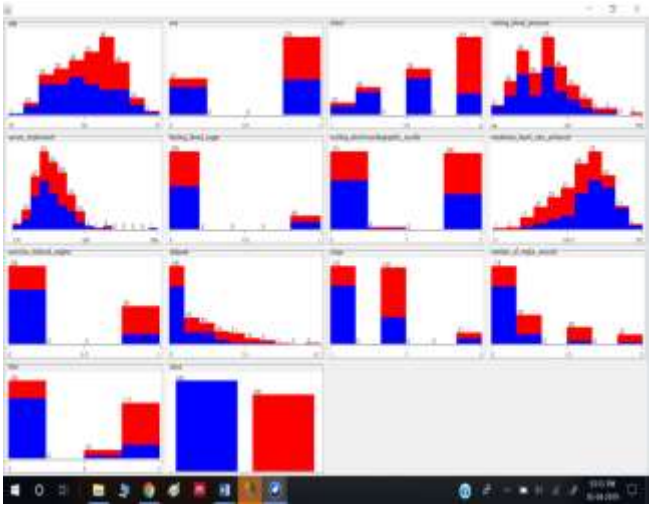

Fig. 3: Frequency Distribution of Cleveland Heart Disease Dataset for all Attributes

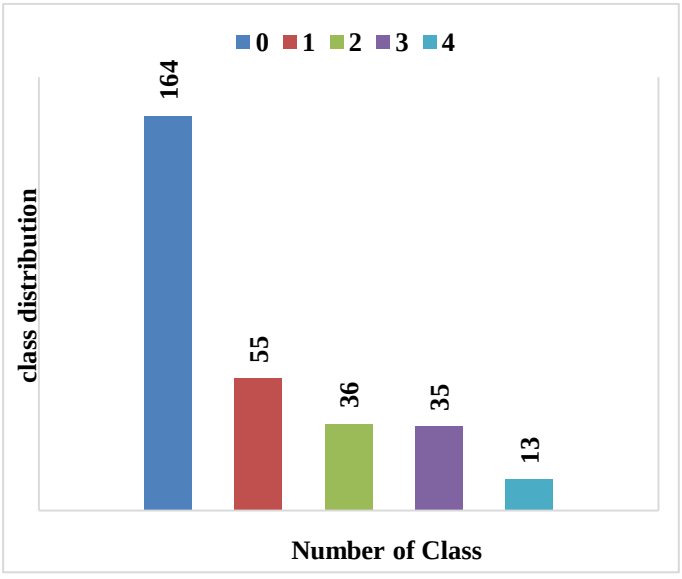

Fig. 4: Class Distributions

3.2 Performance metrics

In examining the proposed GBC classifier over the applied datasets, few measures like precision, sensitivity,

accuracy, specificity, kappa, recall, and F-score are used. To represent the efficiency the projected classifier, it is compared with five other classifiers namely, J48, RT (RT), NBTree, Random Forest (RF) and radial basis function (RBF) Network.

Confusion matrix act as a significant step while evaluating the classification performance. It extracts 2×2 matrix for every classifier depending on the classified output. The confusion matrix of the different classifiers is represented in Table 3 over the given heart-statlog dataset. By employing the values gained using confusion matrix, classifier performance is computed. It is absolute from the table, J48 classifies 88 instances are under the present category, and 119 are under the absent case.RT classifies 89 samples as a present and 117 samples as absent types of heart disease. NBTree classifies 90 samples as a present and 127 samples as absent types of heart disease. Random Forest (RF) classifies 94 samples as a present and 127 samples as absent types of heart disease.RBF classifies the 97 of the instance as present and 97 as absent cases in HD prediction out of the 130 instances. The proposed Gradient boost classifier provides 115 samples under the present type of HD and 142 under the absent type of HD.

Confusion matrix for different classifiers is represented in Table 4 over the given Clevelanddataset. By using the values gained using confusion matrix, classifier performances are computed. It is clear from the table, RT (RT) classifies 125 instances are under present category, and 96 are under the absent case. NBTree classifies 134 samples as a present and 106 samples as absent types of heart disease. J48 classifies 135 samples as a present and 103 samples as absent types of heart disease. RBF classifies the 97 of the instance as present and 130 as absent cases in HD prediction out of the 303 instances. Random Forest (RF) classifies 141 samples as a present and 108 samples as absent types of heart disease. The proposed Gradient boost classifier provides 151 samples under the present type of HD and 137 under the absent type of HD. It also provides minimized counts of false positive and true negative values while comparing with conventional classifiers.

3.3 Performance Evaluation of Heart Statlog Dataset

The graphical representation of precision, sensitivity, accuracy, specificity, kappa, recall, and

F-scoreis given below in Figs. 5-7 and the values are given in Table 5. The performance valuesare offered using percentage.

Table 3: Confusion Matrix of Different Classifiers on Heart Statlog Dataset

Experts	Proposed		J48		RT		RBF		NBTree		Random Forest	
	Present	Absent	Present	Absent	Present	Absent	Present	Absent	Present	Absent	Present	Absent
Present	115	5	88	32	89	31	97	23	90	30	94	26
Absent	8	142	31	119	33	117	20	130	23	127	23	127

Table 4: Confusion Matrix of Different Classifiers on Cleveland Heart Disease Dataset

Experts	Proposed		J48		RT		RBF		NBTree		Random Forest	
	Present	Absent	Present	Absent	Present	Absent	Present	Absent	Present	Absent	Present	Absent
Present	151	13	135	29	125	39	138	26	134	30	141	23
Absent	2	137	36	103	43	96	24	115	33	106	31	108

Table 5: Performance Evaluation of Different Classifiers on Heart Statlog Dataset

Classifiers	Sensitivity	Specificity	Precision	Recall	Accuracy	F-Score	Kappa
Proposed	93.49	96.59	95.83	93.50	95.19	94.65	90.27
J48	73.94	78.81	73.33	73.95	76.66	73.64	52.71
RT	72.95	79.05	74.16	72.95	76.29	73.55	52.08
RBF	82.91	84.97	80.33	82.91	84.07	81.86	67.67
NBTree	79.64	80.89	75.00	79.64	80.37	77.25	60.02
Random Forest	80.34	83.00	78.33	80.34	81.85	79.32	63.16

For precision, J48 classifier gives the poor performance of 73.33%. RT gives the precision rate of 74.16% which is more or less demonstrates a similar rate as J48. RBF classifier outperforms the above mentioned method by attaining the precision rate of 80.33, but it fails to outperform the projected Gradient boost classifier which attains the maximum precision rate of 95.83% which shows that it is the better method using precision rate.

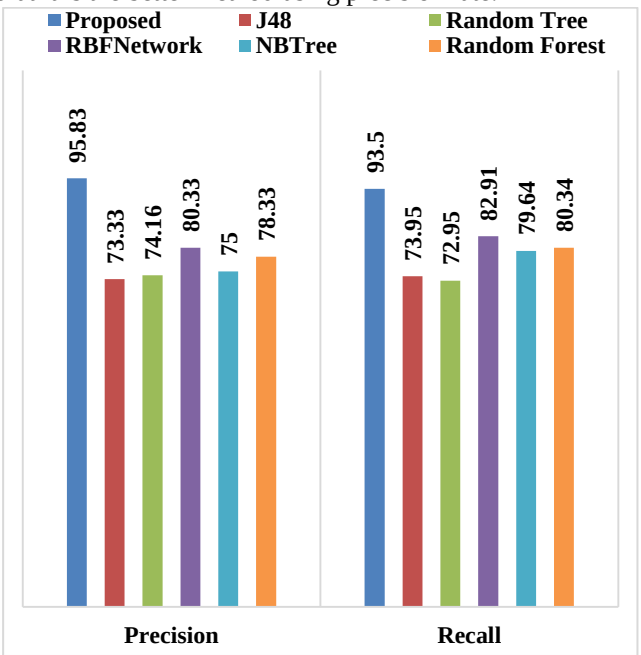

Fig. 5: Comparison of Different Classifiers on Heart Statlog Dataset in terms of Precision, Recall

For recall, as similar to precision, RT and J48 demonstrate more or less the similar recall rate of 72.95% and 73.95%. However, RBF classifier outperforms by obtaining the recall rate of 82.91% which is higher than the other two methods. The projected model attains 93.50% of recall rate when classifying the HD dataset.

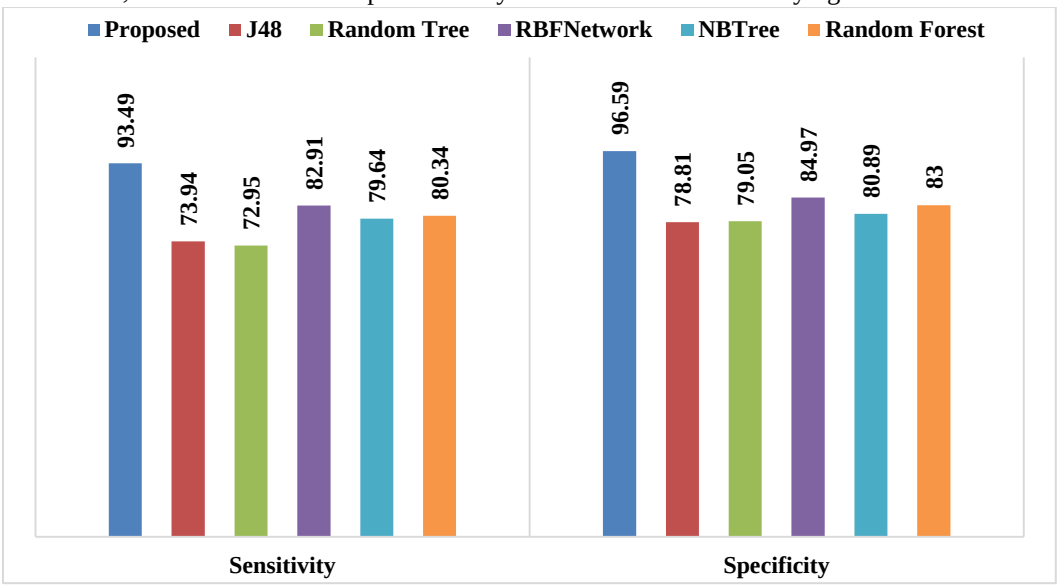

Fig. 6: Comparison of Different Classifiers on Heart Statlog Dataset in terms of Sensitivity, Specificity

Fig.5 demonstrates the comparison among various classifiers over classifying result of the dataset Statlog using Sensitivity, Specificity. For sensitivity, RT classifier gives the poor performance of 72.95%. J48 gives the sensitivity rate of 73.94% which is more or less demonstrates a similar rate as RT. NBTree classifier outperforms the above mentioned method by attaining the sensitivity rate of 79.64, but it fails to outperform the projected Gradient boost classifier which attains the maximum sensitivity rate of 93.45% which shows that it is the better method using sensitivity rate. For specificity, as similar to sensitivity, RT and J48 demonstrate more or less the similar specificity rate of 79.05% and 78.81%. However, Random forest classifier outperforms by obtaining the specificity rate of 83.00% which is higher than the other two methods except for the proposed method. The projected model attains 96.59% of specificity rate when classifying the HD dataset.

For F-score, RT and J48 demonstrate more or less the similar rate of F-score of 73.55 and 73.64 respectively. The classifier RBF attains 81.86% of F-score rate. Above all, the proposed model attains the F-score rate of 94.65 which is best among all. Fig. 6 shows the classifier performance using accuracy for the given HD dataset. RT and J48 demonstrate more or less the similar rate of accuracy of 76.29 and 76.66 respectively. The classifier RBF attains 84.07% of accuracy rate. Above all, the proposed model attains the accuracy rate of 95.19 which is best among the compared methods.

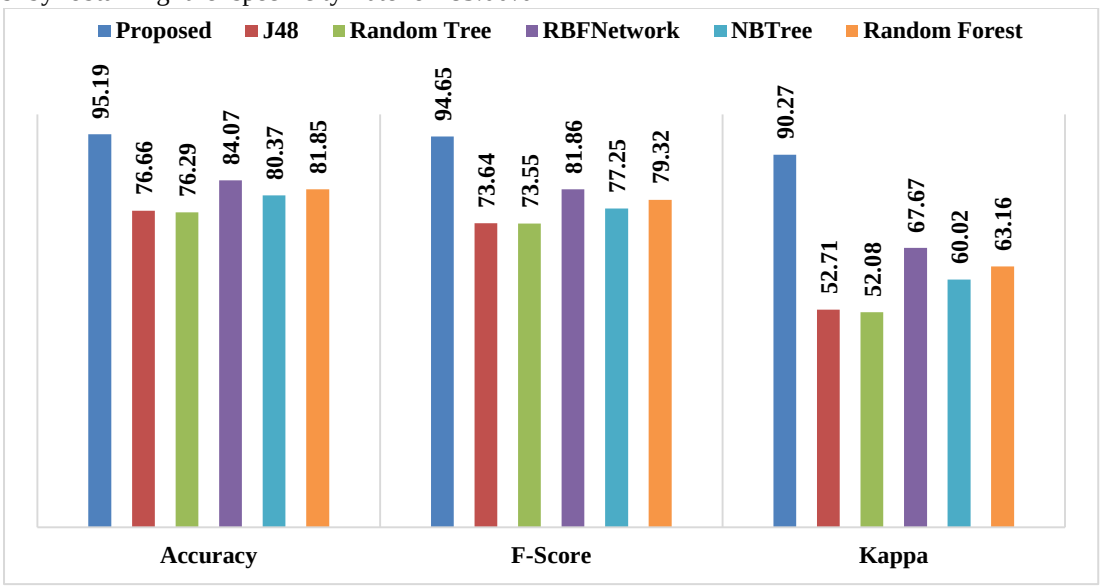

Fig. 7: Comparison of Different Classifiers on Heart Statlog Dataset in terms of Accuracy, F-score, Kappa

Fig. 6 shows the classifier performance using Kappa Value for the given HD dataset. For kappa-value, RT gives the poor performance of 52.08%, and J48 gives the Kappa Value rate of 52.71%. RBF classifier outperforms the above mentioned method by attaining the Kappa Value of 67.67%, but it fails to outperform the projected Gradient boost classifier which attains the maximum Kappa Value of 90.27%. Therefore, for the given HD dataset, the proposed method attains the enhanced performance for all metrics like precision, sensitivity, accuracy, kappa, specificity, recall, and F-score.

3.4. Performance Evaluation of Cleveland Dataset

The graphical representation of precision, sensitivity, accuracy, specificity, kappa, recall, and F-score is given below in Figs. 8-10 and the values are given in Table 6. The performance values are offered using percentage.

Table 6: Performance Evaluation of Different Classifiers on Cleveland Heart Disease Dataset

Classifiers	Sensitivity	Specificity	Precision	Recall	Accuracy	F-Score	Kappa
Proposed	98.69	91.33	92.07	98.69	95.05	94.81	90.09
J48	78.94	78.03	82.32	78.94	78.55	80.59	56.64
RT	74.41	71.11	76.21	74.40	72.94	75.30	45.38
RBF	85.19	81.56	84.15	85.18	83.49	84.66	66.81
NBTree	80.24	77.94	81.71	80.24	79.21	80.97	58.06
Random Forest	81.98	82.44	85.98	81.97	82.18	83.93	63.95

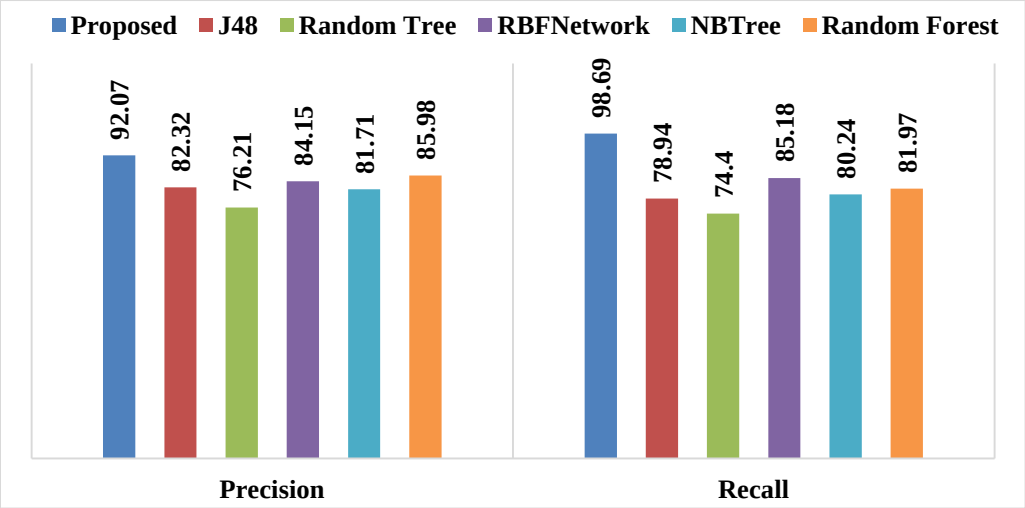

Fig. 8: Comparison of Different Classifiers on Cleveland Heart Disease Dataset in terms of Precision, Recall

For precision, RT classifier gives the poor performance of 76.21%. NBTree gives the precision rate of 81.71%. RF classifier outperforms the above mentioned method by attaining the precision rate of 85.98, but it fails to outperform the projected GBC which attains the maximum precision rate of 92.07% which shows that it is the better method using precision rate. For recall, as similar to precision, RT and NBTree demonstrate more or less the similar recall rate of 74.40% and 80.24%. However, RBF classifier outperforms by obtaining the recall rate of 85.18% which is higher than the other two methods. The projected model attains 98.69% of recall rate when classifying the HD dataset.

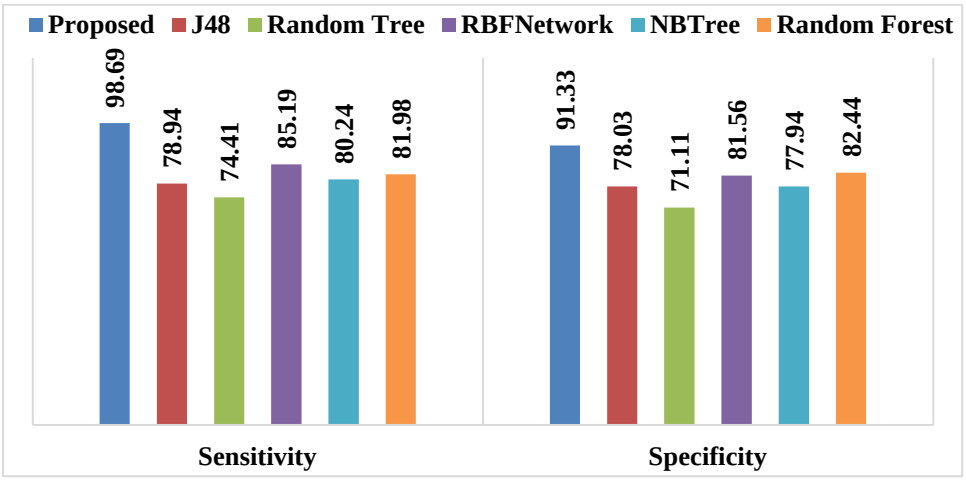

Fig. 9: Comparison of Different Classifiers on Cleveland Heart Disease Dataset in terms of Sensitivity, Specificity

Fig.9 demonstrates the comparison among various classifiers over classifying result of the dataset Clevelandusing Sensitivity, Specificity. For sensitivity, RT classifier gives the poor performance of 74.41%. J48 gives a sensitivity rate of 78.94%. RBF classifier outperforms the above mentioned method by attaining the sensitivity rate of 85.19, but it fails to outperform the projected Gradient boost classifier which attains the maximum sensitivity rate of 98.69% which shows that it is the better method using sensitivity rate. For specificity, as similar to sensitivity, NBTree and J48 demonstrate more or less the similar specificity rate of 77.94% and 78.03%. However, Random forest classifier outperforms by obtaining the specificity rate of 82.44% which is higher than the other two methods except for the proposed method. The projected model attains 91.33% of specificity rate when classifying the HD dataset.

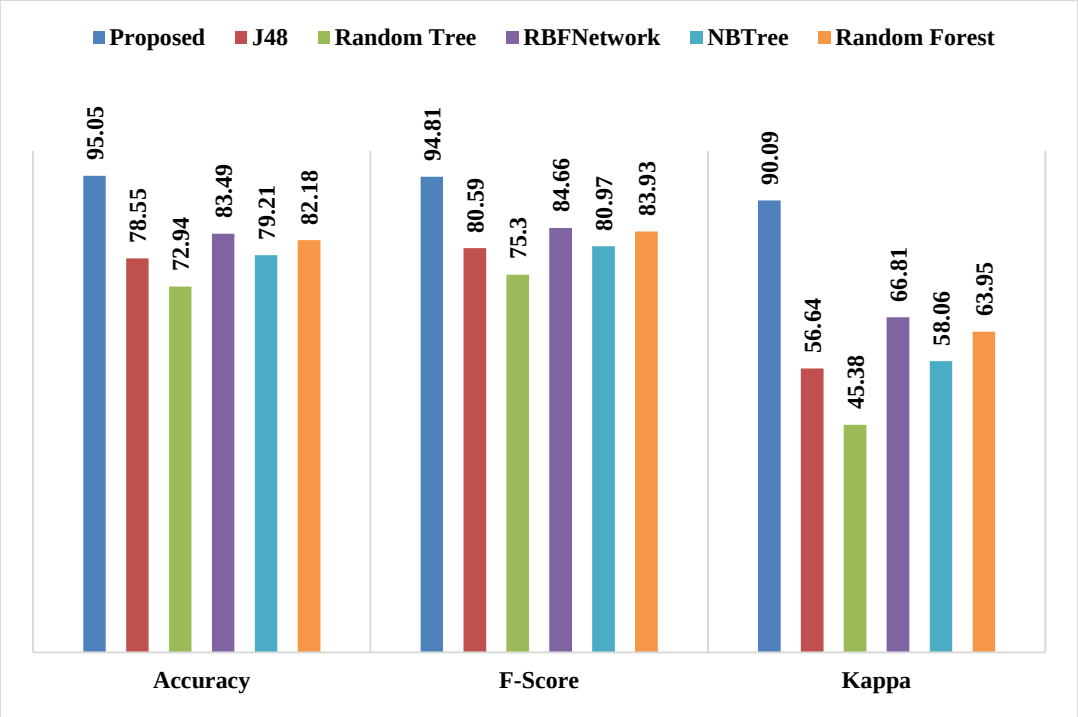

Fig. 10: Comparison of Different Classifiers on Cleveland Heart Disease Dataset in terms of Accuracy, F-score, Kappa

For F-score, NBTree and J48 demonstrate more or less the similar rate of F-score of 80.97 and 80.59 respectively. The classifier RBF attains 84.66% of F-score rate. Above all, the proposed model attains the F-score rate of 94.81 which is best among all.

Fig. 10 shows the classifier performance using accuracy for the given HD dataset. NBTree and J48 demonstrate more or less the similar rate of accuracy of 79.21 and 78.55 respectively.

The classifier RBF attains 83.49% of accuracy rate. Above all, the proposed model attains the accuracy rate of 95.05 which is best among the compared methods.

Fig. 10also shows the classifier performance utilizing Kappa Value for the given HD dataset. For kappa-value, RT gives the poor performance of 45.38%, and J48 gives the Kappa Value rate of 56.64%.

RBF classifier outperforms the above mentioned method by attaining the Kappa Value of 66.81%, but it fails to outperform the projected Gradient boost classifier which attains the maximum Kappa Value of 90.09%.

Therefore, for the given HD dataset, the proposed method attains the enhanced performance for all metrics like precision, sensitivity, accuracy, kappa, specificity, recall, and F-score.

Table 7: Comparison with Recent Methods for Heart Disease Dataset in terms of Accuracy

Classifiers	Accuracy
The Proposed Model	**95.19**
J48	76.66
RT	76.29
RBF	84.07
NBTree	80.37
Random Forest	81.85
Rotation Forest Ensemble Classifiers(2001)	80.49
Khemphila and Boonjing (2011)	80.99
Shouman et al. (2011)	84.10
Pruned J48 DT (2013)	73.79
Chaurasia and Pal (2013)	83.49
Subanya and Rajalaxmi (2014)	86.76
Nahar et al. (2014)	69.11
Extreme Learning Machine (2015)	80.00
Hybrid Genetic Fuzzy Model (2015)	86.00
Ismaeel et al. (2015)	86.50
El-Bialy et al. (2015)	78.54
LR (2016)	85.00
Paul et al. (2016)	80.00
Verma et al. (2016)	80.68
Vivekanandan et al. (2017)	83.00
Ensemble Model (2019)	88.88
Amin et al. (2019)	87.41

Table 7 gives a comparison of the proposed method with the recently projected classifier using accuracy for the given datasets. The pictorial representation is given in Fig. 11. The classifiers such as J48, RT, RBF, Rotation Forest Ensemble Classifiers, Pruned J48 DT, Extreme Learning Machine, Hybrid Genetic Fuzzy Model, LR and Ensemble Model are compared with the projected Gradient boost classifier. It is clear from the table that the classifier Pruned J48 DT is the worst among all by attaining the accuracy rate of 73.79. Subsequently, the classifiers J48 and RT are the next worst performer which obtains more or less the same accuracy rate of 76.66 and 76.29. Similarly, the classifiers Rotation Forest Ensemble Classifiers and Extreme Learning Machine are the next worst performer which obtains more or less the same accuracy rate of 80.49 and 80.00. The classifiers RBF, LR, Hybrid Genetic Fuzzy Model, and Ensemble Model exhibit the accuracy rate of 84.07, 85.00, 86.00 and 88.88 respectively. The other methods do not give a notable performance. The Ensemble Model is the classifier which outperforms all the other classifiers, but it fails in works well when compared to projected Gradient boost classifier. Therefore, the projected method is superior by gaining the accuracy rate of 95.19 to classify the given HD dataset.

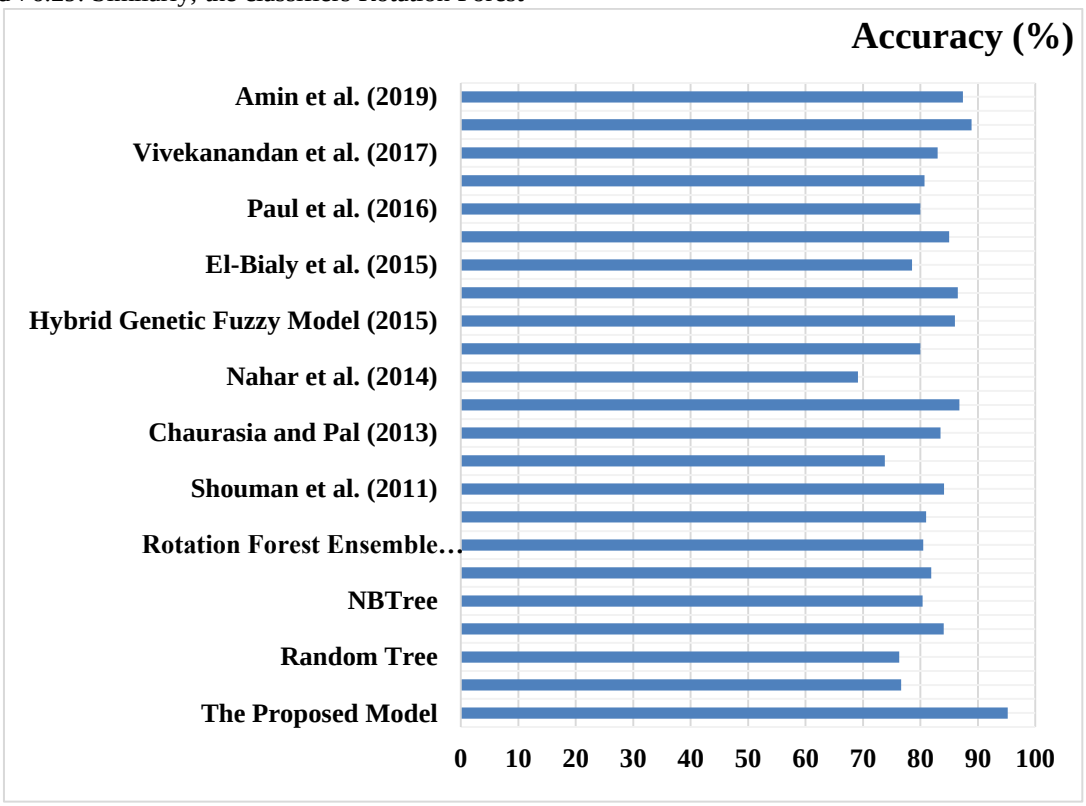

Fig. 11: Comparison of Recently Proposed Method with Our Method

IV. CONCLUSION

Various methods of data mining were aimed at HD and have been used to distinguish the HD patients to identify various HD sub-kinds and to compute HD severity. Though different classifier models have been presented, there is still a need to enhance the performance. In this paper, the GBC model to predict the HD efficiently. To examine the proposed GBC for the HD prediction, numerous analyses were performed and the results of the dataset. For the performance validation of GBC, two benchmark datasets namely Heart-Statlog and Cleveland were used. The projected method is superior to other methods on both the applied dataset.

REFERENCES

1. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, et al. ESC guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J 2016;2015(ehw128). http://dx.doi.org/10.1093/eurheartj/ehw128.
2. Aljaaf AJ, Al-Jumeily D, Hussain AJ, Dawson T, Fergus P, Al-Jumaily M. Predicting the likelihood of heart failure with a multi level risk assessment using a decision tree. Third international conference on technological advances in electrical, Beirut, Lebanon; 2015.
3. Cowie MR. The heart failure epidemic. Medicographia 2012.
4. Son C-S, Kim Y-N, Kim H-S, Park H-S, Kim M-S. Decision-making model for early diagnosis of congestive heart failure using rough set and decision tree approaches. J Biomed Inform 2012;45:999–1008. http://dx.doi.org/10.1016/ j.jbi.2012.04.013.
5. Roger VL. The heart failure epidemic. Int J Environ Res Public Health 2010;7: 1807–30.
6. Dickstein K, Cohen-Solal A, Filippatos G, McMurray JJV, Ponikowski P, Poole-Wilson PA, et al. ESC guidelines for the diagnosis and treatment of acute and chronic heart failure 2008 the task force for the diagnosis and treatment of acute and chronic heart failure 2008 of the European Society of Cardiology. Developed in collaboration with the heart failure association of the ESC (HFA) and endorsed by the European Society of Intensive Care Medicine (ESICM). Eur Heart J 2008;29:2388–442.
7. Yang G, Ren Y, Pan Q, Ning G, Gong S, Cai G, et al. A heart failure diagnosis model based on support vector machine. 2010 3rd international conference on biomedical engineering

and informatics (BMEI), vol. 3; 2010. p. 1105–8.
8. Gharehchopogh FS, Khalifelu ZA. Neural network application in diagnosis of the patient: a case study, Abbottabad; 2011.
9. Masetic Z, Subasi A. Congestive heart failure detection using random forest classifier. Comput Methods Programs Biomed 2016;130:54–64.
10. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet components of a new research resource for complex physiologic signals. Circulation 2000;101:e215–20. http://dx.doi.org/10.1161/01.CIR.101.23.e215.
11. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care 2010;48:S106–13. http://dx.doi.org/10.1097/MLR.0b013e3181de9e17.
12. Zheng Y, Guo X, Qin J, Xiao S. Computer-assisted diagnosis for chronic heart failure by the analysis of their cardiac reserve and heart sound characteristics. Comput Methods Programs Biomed 2015;I22:372–83.
13. Tobin. Estimation of relationships for limited dependent variables. Econometrica: journal of the Econometric Society, pages 24–36, 1958.
14. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acmsigkdd international conference on knowledge discovery and data mining, pages 785–794. ACM, 2016
15. Friedman, T. Hastie, and R. Tibshirani. The elements of statistical learning, volume 1. Springer series in statistics New York, 2001
16. http://archive.ics.uci.edu/ml/datasets/statlog+(heart)
17. http://archive.ics.uci.edu/ml/datasets/heart+disease