Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

A Multi-Stage Approach Combining Feature Selection with Machine Learning Techniques for Higher Prediction Reliability and Accuracy in Heart Disease Diagnosis

Avijit Kumar Chaudhuri[1]
[1] Research Scholar, Dept of Computer Application, Seacom Skills University, Kendradangal, Bolpur, Birbhum, PIN-731236, West Bengal, India

Dilip K. Banerjee[2]
[2] Professor, Dept of Computer Application, Seacom Skills University, Kendradangal, Bolpur, Birbhum, PIN-731236, West Bengal, India

Anirban Das[3]
[3]University of Engineering & Management, Kolkata

Arkadip Ray[4]
[4]Dept of Information Technology, Govt. College of Engineering & Ceramic Technology, Kolkata, Pin-700010, West Bengal, India

***Abstract*—The accuracy in predicting Heart Disease (HD), obtained using different data mining approaches, is around 85 percent so far. An error of 15 percent or so is either type 1 or type 2, meaning a person with HD goes un-detected (type 1), or a person without HD undergoes an HD treatment (type 2). Several studies exist on the application of ensemble techniques; however, the quest for increasing the accuracy levels will continue until an approach provides 100 percent accuracy and reliability. The literature review shows that there are no single data mining techniques that give consistent results for all types of healthcare datasets, and the performance of data mining techniques depends on the type of dataset [1-3]. In this paper, the authors compared various classification methods and feature selection techniques on the same dataset. In this paper the authors compared the performance of various classification methods and feature selection techniques on the same dataset.**

This paper combines two stages, first identifying the significant features using different methods and then testing the change in accuracy of prediction using different standalone Machine Learning (ML) algorithms like Naive Bayes (NB), Support Vector Machines (SVM), Logistic Regression (LR), Decision Tree (DT)) and Ensemble Classifiers like Random Forest (RF), Gradient Boosting (GDB) and Extra Trees (ET) for predicting heart disease (HD). A survey of patients shows that the significant factors considered, by doctors, as the first step to assess the possibility of heart disease include age, sex, test results of blood pressure, cholesterol, and echo-cardio-gram (ECG). This study shows that the prediction accuracy using these features can produce accuracy only to 51 percent. Feature selection techniques may be used to recognize and delete unwanted, obsolete, and redundant attributes from dataset that do not contribute to the accuracy of the predictive model which can potentially reduce the accuracy of the model.

Fewer attributes are advantageous because they minimize the complexity of the model, and a simpler model is simpler to understand and explain. Most of the research article explains the estimation of heart disease in the medical profession by the use of data science. As several research studies have done on this issue, however, the accuracy of forecasts still needs to be improved. Thus, this study focuses on feature selection methods, and ML algorithms are used for the analysis of observations and the enhancement of accuracy. By this research, authors achieved an accuracy of 96.77 % and a precision level of 100%.

Keywords*—Multistage approach; Heart Disease; Feature Selection; Machine Learning Techniques; Type 1 & Type 2 Errors; Prediction reliability and accuracy.

I. INTRODUCTION

Ensemble methods provide a way to increase accuracy and reliability by combining a set of base classifiers. The basic goal of the ensemble models is to reduce variance in the model and to increase the precision of the predictions. So, one of the most important research questions in this study is whether ensemble models will still increase the precision of disease predictions? Despite their high-predictive efficiency, for two key reasons, other models may be selected over ensemble models. First, predictions typically take a long time for large ensembles, since multiple inducers are used to combine a single prediction [4].

This problem continues to be important for predictive real-time systems [5]. Moreover, it is almost impossible to perceive the outputs of the ensemble since they consist of the outputs of several inducers. This property typically forbids the use of Ensemble Models in domains that involve a simple and logical interpretation of individual decisions (e.g., medicine, insurance, etc). Many studies related to the application of ensemble techniques [6]; however, the quest for increasing the accuracy levels will continue until an approach provides 100 percent accuracy and reliability. The accuracy for predicting diseases such as heart disease (HD) is of utmost importance as errors result in fatalities. So far, studies on the prediction of heart disease show accuracy to the tune of 85%.

Heart disease (HD) is a disease of concern as it results in restriction in activities, an increase in expenditure, mortality, or becomes a risk factor to fatalities caused by other diseases such as COVID-19. Hence, there is a need for very high precision in the prediction of HD. Several researchers have suggested the use of machine learning techniques (MLT) for the prediction of HD. However, results show an error rate of 15 percent or so in most of the approaches. Further, this error is either type 1 or type 2, meaning a person with HD goes un-detected (type 1), or a person without HD undergoes an HD treatment (type 2). This incorrect diagnosis happens due to incomplete medical testing.

An attempt to seek doctors' opinions regarding HD's significant features did not yield the desired result as around 15 of them opined that the features depend on patients' study. The authors interviewed around 150 patients in India to identify doctors' tests for judging probable HD. In 90 percent cases, the authors found that doctors inquired about blood pressure, cholesterol, blood sugar, and age and sex to arrive at a prescription. The rest of the tests, namely, fluoroscopy, Holter monitor, and echocardiogram, are suggested if patients suffer from chest pain or chest angina or any discomfort.

HD Data Set, available in UCI machine learning repository, records 13 features of a patient. Earlier researchers showed 85 percent accuracy in prediction using MLT.

Researchers have shown that the accuracy of classification decreases with an increase in the number of features [7,8] as shown in Figure 1. Feature selection is a way to remove the unnecessary features that are considered to be non-essential to the data mining task.

The objective of the Feature Selection is to examine the appreciable arrangement of features. A significant pre-processing step for classification is the feature subset selection. In biology, where a large number of features describe structures or processes, the elimination of irrelevant and redundant information has a number of advantages over a reasonable amount of time. It allows the classification system, with a limited subset of features, to achieve good or even better solutions, allows for faster classification, and helps the human expert to concentrate on a relevant subset of features, thus providing useful biological knowledge [9,10].

The literature review shows that there are no single data mining techniques that give consistent results for all types of healthcare datasets, and the performance of data mining techniques depends on the type of dataset [1-3]. In this paper, the authors compared various classification methods and feature selection techniques on the same dataset. The authors compared the performance of various classification methods and feature selection techniques on the same dataset.

The authors, in this paper, suggest a two-stage prediction process. First, identify the most significant features; second, predict using the crucial features. The authors used four practical filter approaches: the Information-gain, Relief-F, One-R, and Gain-ratio.

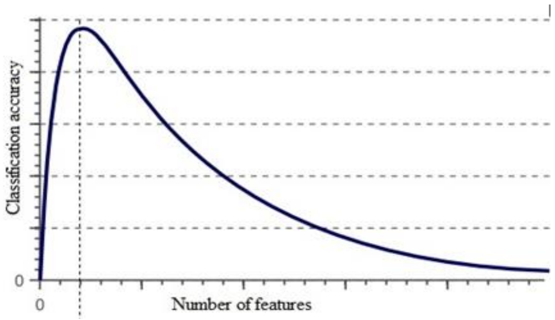

Fig. 1. Classifier Efficiency Relationship with Dimensionality

II. PERFORMANCE OF MLT AND NEED FOR FS

Previous studies concluded that different MLT have different accuracy levels, and none of them is proved superior. Thus, there is a need for comparative performance evaluation amongst the widely used MLTs.

Machine learning (ML) and feature selection (FS) techniques can help make better decisions and classify many diseases with accuracy levels, but not all FS techniques contribute equally to classification [11]. Some have issues of overfitting or dependence on classifier while some methods ignore classifiers and feature dependencies. Filter methods (FM) are free from bias associated with learning models as in case of wrapper or embedded models. Besides, FM are straightforward, simple, fast and widely used [12]. In multi-variate FM techniques, it is capable to model feature dependencies and its complexity is lower than wrapper methods (WM). WM has the risk of overfitting and sticking to local minima, while embedded method's success depends on choice of classifier. In general, two challenges are associated with bioinformatics – large input dimensionality and small sample sizes [13]. In this paper, the filter methods were applied to enhance predictability. Results from four methods - Relief-F, Information gain, Gain Ratio, and One-R techniques were used for classification and compared to suggest the best combination of FS and MLT.

Heart disease (HD) is a common disease and cuts across ages WHO reported HD as the number one cause of death globally. Several studies [14,15] demonstrated ML techniques for HD classification. Varatharajan et al. [16] suggested a genetic algorithm (GA) with binary particle swarm optimization (BPSO) for the prediction of coronary artery disease. Researchers also recommended various classification and regression methods to identify the hidden values and useful data from the healthcare dataset [17]. These studies' classification accuracy is shown in table 1, indicating a maximum of 84.14% accuracy levels. Most of these studies mainly concentrate on accuracy calculations. Sensitivity, specificity, predictive values, and likelihood ratios (LRs) are all common forms of describing test results. Receiver Operating Characteristic (ROC) curves measure sensitivity versus specificity over a spectrum of values with the potential to forecast dichotomous outcomes. A further indicator of test success is the area under the ROC curve. High sensitivity correlates to strong negative predictive performance and is the "rule-out" study's optimal property. High specificity correlates to a strong positive predictive value and is the optimal property for the "rule-in" study [18].

Cohen’s Kappa statistic is an inter-rater cooperation indicator for categorical variables. This statistic determines the degree to which two tests agree with diagnostic categorization. It is commonly considered as a more rigorous indicator than a simple percent accuracy measurement because Kappa takes into consideration an arrangement that happens by chance [19].

Table 1 shows a comparison of performance of different machine learning on medical databases. The results do not include the significant statistical measures namely, Precision,

Kappa and ROC/AUC. Thus, higher prediction accuracy determined by previous research only leads to partial conclusion. The table also indicate that most of research have focused on ML classification strategies without minimizing the irrelevant features.

The approach suggested in this paper substantiates the findings with the Precision, Kappa and ROC/AUC statistics and shows that feature selection enables improved results. Feature selection techniques enables recognize and delete unwanted, obsolete, and redundant attributes from data that do not contribute to the accuracy of the predictive model which can potentially reduce the accuracy of the model [20,21]. Fewer attributes are advantageous because they minimize the complexity of the model, and a simpler model is simpler to understand and explain.

TABLE I. COMPARISON OF THE STUDIES USING UCI HEART DISEASE DATASET

Year	**Method**	**Classification Accuracy (%)**	**Sensitivity/ Specificity**	**Precision**	**ROC/AUC Area**	**Kappa Statistic**	**Features Selected**
[22]		84% (around)	×	×	×	×	13
[23]	Bagging Algorithm, J48 DT	Bagging (81.41%) J48 DT (78.90%)	Bagging (74.93/ 86.64) J48 DT (72.01/ 84.48)	×	×	×	
[24]	One Dependency Augmented Naïve Bayes Classifier (ODANB), NB	ODANB (80.46%) Naive Bayes (84.14%)	×	×	×	×	
[17]	Binary Particle Swarm Optimization (BPSO) and Genetic Algorithm (GA)	81.46	×	×	×	×	
[15]	Cost-Sensitive Case-Based Reasoning (CSCBR)	×	0.90/0.87	×	×	×	
[14]	Artificial Neural Networks	70%	×	×	×	×	
[25]	Weighted Fuzzy Rules	62.35%	0.766/0.766	×	×	×	
[26]	Rotation Forest, Levenberg-Marquardt	91.20%	✓	×	✓	×	
[27]	Without Voting K-Nearest Neighbor	97.10%	0.935/0.987	×	×	×	13
[28]	DT	83.90%	81.6/83	×	×	×	
[29]	CART	83.49%	✓	✓		✓	
[30]	NB, DT, J48	NB (83.40%) DT (76.20%) J48 (77.50%)	✓	×	×	×	
[31]	Majority (Vote Based) Ensemble Technique	81.82%	0.7368/0.9286	×	×	×	
[2]	Feature Selection Based Least Square Twin Support Vector Machine (LSTSVM)	85.59%	0.8571/0.9091	×	×	×	11
[32]	J48	56.76%	×	×	×	×	
[33]		91.10%	1/0.84	×	✓	×	
[16]	Dynamic Time Warping (DTW)	×	95.9/94	×	×	×	
[20]	Minimum Redundancy Maximum Relevance Feature Selection (MRMR/FS) LR & SVM	84.85%	×	×	×	×	
Our Study		93.55	0.80/1	1	0.9	0.8	6

Most of the research article explains the estimation of heart disease in the medical profession by the use of data science. As several research studies have done on this issue, however, the accuracy of forecasts still needs to be improved. Thus, this study focuses on feature selection methods, and ML algorithms are used for the analysis of observations and the enhancement of accuracy. In this paper, the authors address three research questions-

First - does all features for predicting HD have equal significance?

Second - does measuring cholesterol, blood pressure, and ECG lead to right prediction?

Third - does ensemble models always improve accuracy?

Some studies [34] show FS algorithms’ use to improve the classification. More generally, the FS algorithms include the selection of correlation-based features and selection of consistency filter-based features [35]. Authors [16] have also proposed a heuristic evaluation approach to evaluate the sub-sets of functions in the correlation-based collection of features. The correlation of ranks of features identified for the training and test period determines the significant features. That is, features with high correlation get assigned as significant features, while the predictive models overlook features with a low-rank correlation [25]. Consistent-filter-based applications propose a selection of significant features based on each feature’s consistency values [16,36-37]. Chaudhuri et al. [38] suggested a stepwise K-means clustering method to minimize errors and showed that all features’ inclusion reduces results’ sensitivity. Most importantly, it showed that the exclusion of cholesterol led to a reduction in type 1 error.

In this paper, the authors discuss the efficacy of the various types of FS methods. The authors propose an algorithmic approach to identify the significant features associated with heart disease. This approach is a two-stage iterative approach. In this approach, first, a random subset of the number of features, from the HD dataset comprising 155 patients available in the UCI Cleveland repository, is chosen for each iteration. Next, the authors apply the feature ranking techniques, namely Relief-F, Information gain, Gain Ratio, and One-R techniques, to determine the significant features. Figure 2 gives the flow chart of the algorithm.

After the selection of significant features, the authors classify and predict, using data on significant features, using seven ML classifiers, namely Naive Bayes (NB), Decision Tree (C4.5), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Extra Trees (ET), and Gradient Boosting (GDB).

The authors note the sensitivity, specificity, ROC curve value, calculated accuracy of training and test samples, and discriminating performances of the mentioned classifiers. It then considers a revised subset of features and compares the results with the previous one. If the revised set shows better results than the original subset, then a next revised subset is chosen, and the authors perform similar iterations to see if the results improve. Thus, the iterations continue until the results show improvements. The authors choose a subset based on the method of inclusion. Initially, the authors apply the classifiers on the dataset comprising the most significant features. Subsequently, it includes the next best feature to see if the performance improves. The authors continue this iteration until accuracy continues to improve. Different FS methods may show different rankings, and so the authors perform the iterations separately for different FS results. Fig. 2 gives the flow chart of the algorithm.

Therefore, this analysis aims to empirically test the efficacy of the classification resulting from the seven ML approaches using four FS techniques for cardiovascular disease risk prediction. This study's findings will provide an insight into the most significant factors a medical practitioner may bear in mind for predicting HD.

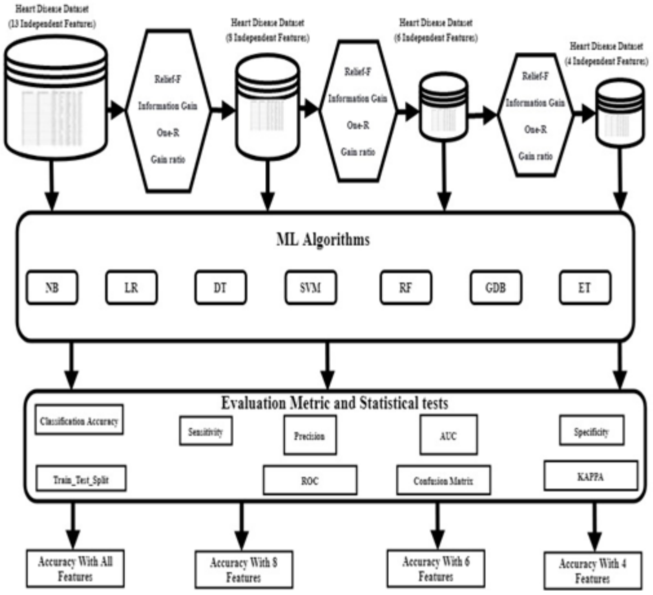

Fig. 2. Flowchart of the algorithm proposed by authors

The authors organize this paper into ten sections. The next section describes the relevant literature in the field of application of machine learning technique (MLT) and feature selection, including its application in the prediction of heart disease. Section II discusses the feature selection techniques and their characteristics. Section III compares classifiers' characteristics, while section IV and V describe the classification algorithm and dataset and its attributes respectively. Section VI describes the classification metrics discussed in the paper and section VII details the significance of features obtained using FS methods. Section VIII shows the computation of HD prediction, and IX discusses the results and findings. The authors conclude their study in the last section, indicating the future scope of work.

III. RELEVANT LITERATURE

Anderson et al. [69], in their study based on a multi-factorial approach on Framingham Heart Study (FHS) data set, experienced that consideration of all the risk factors is probably the best strategy for the prevention of coronary heart disease (CHD). The authors examine precursors for coronary heart disease (CHD) and identified causal risk factors. They concluded that blood cholesterol (especially total and low-density lipoprotein cholesterol), blood pressure, cigarette smoking, diabetes, and unhealthy dietary habits are the best-established risk factors for CHD [39-40]. Greenland et al. [41] analyze the Framingham Heart Study (FHS) data set to find the risk factors for fatal coronary heart disease (CHD) and nonfatal myocardial infarction. They showed total cholesterol of a minimum value of 240 mg/dL (>= 6.22 mmol/L), systolic blood pressure of at least 140 mm Hg, diastolic blood pressure at least 90 mm Hg, cigarette smoking, and diabetes were the crucial factors for CHD.

Prakash et al. [42], in a recent study, introduced Optimality Criterion function selection (OCFS) for extrapolation on HD prediction and clinical diagnosis of HD. Researchers are improving their rough set FS method on information entropy (RFS-IE). This analysis compares the OCFS with the RFS-IE using different data sets within computational time, prediction accuracy, and error rate. Compared with the other method, the OCFS method will take minimum execution time.

Pouriyeh et al. [43] have compared various machine learning techniques and analyzed their findings from a precision perspective. They used a small data set and observed that Support Vector Machine (SVM) as a good classifier. They applied methods - Bagging, Boosting, and Stacking to find the change in accuracy levels. On applying the stacking technique, Multilayer Perceptron (MLP) showed 84.15 percent accuracy, higher in comparison to SVM.

Long et al. [44] used the firefly algorithm in their work on disease prediction and made use of rough set theory to train the classifier. The findings are comparable to other forms of classification, such as Naive Bayes (NB) and SVM. The proposed research surpasses convergence speed, processing time and increases prediction accuracy to 87.2%. The analysis fails in case of a large number of attributes as the rough set attribute becomes unmanageable.

Nahar et al. [45] conducted a study by contrasting various classifiers for HD extraction. They showed that SVM has potential if absolute precision is optimized as the output measure. The experiment's findings revealed that using the medical knowledge motivated feature selection (MFS) technique significantly improved the output for most of the classifiers for most datasets, especially accuracy.

IV. FEATURE SELECTION METHODS

There are several methods for choosing features in the area of machine learning. The primary purpose of such approaches is to remove outdated or redundant features from the dataset. There are two types of FS methods, namely, the wrapper and filter method.

The wrapper checks and selects attributes based upon the target learning algorithm's accuracy estimates. The wrapper effectively checks function space by omitting some characteristics and checking the effect of the elements' omission on the predictive metrics. The position that makes a significant difference in the learning process means that it is essential and is viewed as a high-quality function.

On the other hand, whatever the learning algorithm, "Filter" uses the data's general features and functions. The number correlation between a set of characteristics and the target function is explicitly used by "Filter." The target variable value determines the amount of similarity between the target variable and the characteristics. Filter-based solutions are not classification-based and are typically quicker and more flexible than wrapper-based methods. We also have a low complexity of computation. In filter algorithms, the features are first rated and graded by class mark significance and then picked by a threshold value [46]. In each of these applications, collection algorithms have an assessment value for every feature. This paper uses four significant filter approaches: the Information-gain, Relief-F, One-R, and Gain-ratio. The following sections describe the characteristics of these approaches.

A. *Information Gain*

Information Gain is one of the methods commonly used in a variety of applications for evaluating features. This approach controls all functions according to a user-defined target value. The entropy description for the feature rank is embedded in the cycle of Information Gain. This approach was formulated to estimate each attribute's quality using entropy by calculating the difference between the prior entropy and the post entropy [47].

Information gain (relative entropy or Kullback-Leibler divergence) is a function of the difference between two probability distributions in probability theory and information theory. This approach assesses a feature M by measuring the amount of information obtained from factor N of class (or group), defined as follows:

$$I(M) = H(P(N)) - H\left(P\left(\frac{N}{M}\right)\right) \tag{1}$$

In particular, it tests the difference between the marginal distribution of measurable N on the assumption that it is independent of $H(P(N))$ and the conditional distribution of N on the assumption that it is dependent on $H\left(P\left(\frac{N}{M}\right)\right)$. If M is not expressed differently, N will be independent of M, which means that M will have limited benefit value for data and vice versa.

B. *Relief-F*

Kira & Rendell [48] developed Relief using the distance-based metric method, which weighs each feature depending on its significance (correlation) to the target class. However, reliability is inefficient as it can handle only two-class problems and does not handle redundant features. The updated version of Relief, known as ReliefF [47], can manage multi-class problems and handle incomplete and noisy data sets. Nonetheless, the elimination of redundant functions fails.

The selection of features is an instance-based approach that evaluates a function by extracting samples from different but similar classes and their output. For each feature M of the same class and each of the different classes, Relief-F selects a random sample N of its closest neighbours. Kononenko [47] defines M as the number of weighted variations and the same classification between different classes.

C. *One-R*

One-R is a simple algorithm proposed by Holte. It produces one rule in the training data for each attribute and then selects the rule with the least error. This method considers all numerically valued features continuous and uses a straightforward approach to split the range of values into several intervals of disjoint. It handles missing values by marking a specific attribute as "missing." This system is among the oldest of these. It produces simple, one-feature rules. Although it is a type of minimal classification, it may be useful as a benchmark for other learning schemes to determine a baseline performance [49].

D. *Gain ratio*

Gain ratio (GR) is a shift that reduces the bias in information gain. The GR will take into account the number and size of divisions when selecting an attribute. It corrects information gain by recognizing the inherent knowledge of a split. The intrinsic information is entropy of branch-wide instance distribution (i.e., how much information we need to say what branch an instance belongs to). The attribute value decreases as the information intrinsically increases [50]. Equation 2 gives the way gain-ratio of an attribute is calculated.

$$GR(\text{attribute}) = \frac{\text{Gain}(\text{attribute})}{\text{Intrinsic-Info}(\text{attribute})} \tag{2}$$

V. CLASSIFICATION ALGORITHMS

Different classification algorithms have different strengths and weaknesses. There is no single learning algorithm for supervised learning, which is best suited to all issues. This section summarizes seven state-of-art-algorithms of supervised learning used in this analysis, namely NB, DT, SVM, LR, RF, ET, and GDB.

A. *Naive Bayes (NB)*

NB classifier is Bayes theorem-dependent and deals with primary probabilistic classification with strong independent assumptions. The presence or absence of any particular variable depends, according to this classifier, on the presence or absence of other variables in the dataset [51-52]. NB classifier deals mainly with conditional probabilities. Bayes' theorem used a formula to calculate likelihood by counting the number of values and value combinations in historical data. Bayes' theorem calculates the probability that an occurrence will occur, given the likelihood that another event will occur already. If X is the dependent event based on the event Y, Bayes' theorem can be stated as follows.

The algorithm counts the number of cases in which X and Y coincide, dividing them by the number of cases in which X occurs alone. The benefit of the Naive Bayes classifier is that it needs a small amount of training data to estimate the parameters (means and variances of variables) required for classification. Since independent variables are known, only the variables' variances shall be determined for each class and not the sum. It proposed for classification of multiple class issues.

B. *Decision Tree (DT)*

Among the most popular and widely used data mining algorithms is the J48 algorithm, derived from the ID3 algorithm. J48 may act as a decision tree or a set of classification rules. This approach makes the set of rules comprehensible and, as such, is preferred in many applications. J48 is an algorithm developed by Ross Quinlan for decision trees [53]. J48 is an expansion of earlier Quinlan's ID3 algorithm. The decision trees created by the J48 can be used for classification, sometimes referred to as statistical classifiers. In 2011, Weka machine learning software developers described the J48 algorithm as "the landmark decision-taking tree method which is probably the most widely used machine learning workhorse in practice to date" [46]. J48 creates decision trees, using the information entropy theory, from a set of training data in the same way as ID3.

Training data is a compilation of S_d = s_{d1}, s_{d2}, s_{d3}, s_{d4} … … .. of already classified samples. Each sample s_{di} consists of a m-dimensional vector $x_{1,i}$, $x_{2,i}$, $x_{3,i}$, … … … … $x_{m,i}$, where the x_j represents attribute values or features of the sample, as well as the class in which S_{di} falls. J48 selects the data attribute at each tree node, which most effectively splits its sample set into subsets enriched in either class. The splitting into the sub tree criterion is the average information gain (difference in entropy). The attribute having the highest uniform value of knowledge is selected when making the decision. Then the J48 algorithm recurs on partitioned sub-lists.

C. *Support Vector Machine (SVM)*

An SVM is a term for a group of linked, supervised statistical and computer science learning methods that analyze data and recognize patterns formed for classification and regression analysis by Cortes and Vapnik [54]. SVM has shown strong performance in several areas for implementation. It creates a hyperplane or set of hyperplanes in a high or infinite-dimensional space that can be used to classify, regress, or other functions. SVM's are useful for classifying data.

SVM is also known as a linear classifier [55], which classifies data by finding an ideal hyperplane dividing d-dimensional data into its two classes with a maximum interclass distance. It uses SVM kernel functions to pass data into a higher-dimensional space capable of data isolation [56].

Classification of data from various datasets is the most common function of machine learning techniques. In SVMs, a data point is viewed as an x-dimensional vector (a set of x numbers), and it differentiates such points with a(x−1) dimensional hyperplane.

SVM is a learning machine that tracks vectors and marks each vector by class in high-dimensional space training [57]. Based on the risk minimization concept, SVM aims at reducing error rates [58].

D. *Logistic Regression (LR)*

The LR classification defines the logistic model parameters (a form of binary regression) and is a specific type of linear regression model, though the dichotomous response factors violate the presumption of normality in generalized regression algorithms. The LR model ensures that the linear representation of the available explanatory variables' observed values is a proper function of the fitted probability of the case. LR's main potential is to generate a simple form of probabilistic classification [59].

The drawbacks are that LR cannot adequately address the problems of explanatory variables with non-linear and interactive results. LR is a type of regression used for estimating a binary dependent variable. It uses the maximum likelihood ratio in generating the LR equation to determine the statistical significance of the variables. Equation 3 describes the LR model.

$$\mathrm{prob}\left(K=1\right)=\frac{1}{1+e^{-\left(b_0+b_1a_1+\ldots\ldots+b_ma_m\right)}} \qquad (3)$$

Where prob(K=1) is the probability of existence of the disease and $\beta_0, \beta_1, \ldots\ldots..\beta_m$ are coefficients of regression.

Within the logistic regression method, there is a linear mo del secret. The normal logarithm of the prob(K=1) ratio gives a linear model in a_i to (1–prob(K=1)) as shown in equation 4 and expanded in linear form as shown in equation 5.

$$f\left(a\right)=\ln\left(\frac{\mathrm{prob}\left(k=1\right)}{1-\mathrm{prob}\left(k=1\right)}\right) \qquad (4)$$

$$=b_0+b_1a_1+\ldots\ldots+b_ma_m \qquad (5)$$

LR is useful in situations where one can predict the presence or absence of a dependent feature or outcome based on independent feature set values. It is similar to a linear regression variant but is suitable for models with a dichotomous dependent variable.

E. *Random Forest (RF)*

RF (or Bagged DTs) is a machine learning ensemble method involving the construction (growth) of multiple DTs via the aggregation (bagging) of bootstraps. Each tree predicts the outcomes and RF uses the method of voting to arrive at the final prediction [60]. This weighted multi-DT vote is usually less risky than a single DT outcome and less susceptible to outliers, reducing uncertainty due to limited facts and reliable predictions [61]. RF has an efficient filtering system and can handle several input parameters without deleting any reduced dimensionality parameters.

RF calculates factor significance scores by measuring the difference in prediction error using the permutation test concept. This score is calculated for each constituent tree, compounded across the whole group, and divided by the standard deviation. This approach combines with a random selection of features to create guided variations in decision trees. The duty for forest development rests with the trees themselves.

F. Extra Trees (ET)

A variant of a random forest is an “extra trees” classifier (ETC), also known as an “Extremely randomized trees” classifier. The entire sample is used at each step in an ET, unlike a RF, and the decision boundary is randomly chosen rather than the best. In real-world cases, performance is comparable to an ordinary, RF, sometimes a bit better [62,63]. In particular, it is an ensemble of decision trees and is associated with other algorithms for decision trees, such as bootstrap aggregation (bagging) and RF. By generating a large number of un-pruned decision trees from the training dataset, the ET algorithm operates. In the case of regression, predictions are made by averaging the prediction of the decision trees or by utilising the majority vote in the case of classification.

G. Gradient Boosting (GDB)

One of the best-supervised machine learning algorithms for regression and classification problems is GDB. The GDB algorithm is likely to make decisions in the form of an ensemble of weak prediction models. It builds a model in a step-by-step fashion as do other boosting approaches, which generalizes them by facilitating the optimization of an arbitrary differentiable loss function [64, 65]. Trees are added to the ensemble model one at a time and are adapted to correct the prediction mistakes of prior models. This is a type of model of a machine learning ensemble known as boosting. For optimization of the arbitrary differentiable loss function and gradient descent, models fit with any algorithm. This gives the technique its name, “gradient boosting,” as the gradient of loss is minimised as the model fits, much like a neural network.

VI. HD DATASET AND ATTRIBUTES

The authors selected the dataset of 155 patients who went for medical examinations from the UCI datasets. TABLE II shows the biometric data collected during the physical examination. Out of 155 patient records, 66 patients have HD, and 89 patients do not have HD.

TABLE II. DESCRIPTION OF DATASET

	Attributes	Description	Values
1	AGE	Age in years	Continuous; Minimum age is 34 and maximum is 74
2	SEX	Male or Female	1=Male, 0 = Female
3	CP	Chest Pain Type	1=typical type, 2= atypical type angina, 3=non-angina pain, 4 = asymptomatic pain
4	TRESTBPS	Rest blood Pressure	Continuous value in mm hg Minimum blood pressure is 94 and maximum is 200
5	CHOL	Serum Cholesterol	Continuous value in mm/dl; Minimum cholesterol is 126 and maximum is 564
6	FBS	Fasting Blood Sugar > 120 mg/dl	1=true, 0=false
7	RESTECG	Resting Electrocardiographic Results	0=normal,1=having ST-T wave abnormality,2=showing probable or definite left ventricular hypertrophy
8	THALACH	Maximum Heart Rate Achieved	Continuous; Minimum value obtained is 71 and maximum is 195
9	EXANG	Exercise Induced Angina	1=true, 0=false
10	OLDPEAK	ST Depression induced by exercise relative to rest	Continuous
11	SLOPE	Slope of the peak exercise ST segment	1=unsloping, 2=flat, 3=down-sloping
12	CA (check for arterial blockages)	Number of major vessels coloured by Fluoroscopy	0 - 3
13	THAL (thalassemia)	Defect Type	3=normal, 6=fixed defect, 7=reversable defect
14	NUM	Diagnosis of HD	0 = no cardiac disorder (<50% narrowing) 1 = has cardiac disorder (>50% narrowing)

VII. CLASSIFICATION METRICS

The authors used the HD dataset applied four filter-based FS approaches, determined the selected features with higher gain, and classified using seven ML techniques to determine the best classification for predicting disease risk. The authors select aggregation algorithms for all features, i.e., 13, followed by 10, 9, 8, 6, and 4 significant features. They use metrics, namely precision and Recall, where equation 6 gives the expression for precision metric.

$$\text{Precision} = \frac{\text{True Positives}\left(T_P\right)}{\text{True Positives}\left(T_P\right) + \text{True Negatives}\left(T_N\right)} \qquad (6)$$

The term precision means the measurements are limited to two values or more. The precision metric determines whether the measurement is correct or reproducible; in other words, the percentage of tests listed as healthy provides precision shown in equation 7.

$$\text{Recall} = \frac{\text{True Positives}\left(T_P\right)}{\text{True Positives}\left(T_P\right) + \text{False Negatives}\left(F_N\right)} \qquad (7)$$

The metric Recall is a proportion of positive factors expected to be positive.

A. Sensitivity and Specificity Analysis

For all health care, diagnosis and procedures are imprecise and subject to error rates. The standard method for evaluating this medical error is a sensitivity and specificity test. A diagnosis and a medical examination should be differentiated. A test is one of many terms used to describe the medical condition; diagnosis is a mixture of numerous tests and observations that demonstrate the patient's pathophysiology. Sensitivity/specificity assessment is necessary for testing and diagnosis are shown in equation 8 and 9. There are test outcomes and an objective indicator of truth or theory in the medical dataset and analysis of sensitivity and specificity. The consistency of test checks done by if-then rules governs the similarity of a new test result to the expected value. The best fit rule specifies the class composition of the study when defining an example of a test.

True Positive (T_P) specifies the number of correct positive predictions (classifications); True Negative (T_N) indicates the number of correct negative predictions; False Positive (F_P) represents the number of incorrect positive predictions; False Negative (F_N) identifies the number of incorrect negative predictions.

The two measures are:

$$\text{sensitivity} = \text{recall} = \frac{\text{True Positives}\left(T_P\right)}{\text{True Positives}\left(T_P\right) + \text{False Negatives}\left(F_N\right)} \quad (8)$$

$$\text{specificity} = \frac{\text{True Negatives}\left(T_N\right)}{\text{False Positives}\left(F_P\right) + \text{True Negatives}\left(T_N\right)} \quad (9)$$

Sensitivity measures a test's tendency to be positive when the condition is present, or how many positive examples of tests are remembered [66]. The sensitivity measures, in other words, test how often anyone finds what they are looking. It falls within several near-synonyms: false-negative rate, recall, type I error, type II error, omission error, or alternative hypothesis. Specificity tests a test's tendency to be negative if the condition does not exist, or how many negative test instances are omitted.

In other words, the specificity tests how often they are searching for what anyone can find. It falls within various near-synonyms: false-positive rate, accuracy, type I error, type II error, commission error, or null hypothesis. Predictive precision provides an overall assessment. Only findings that give high values for all three measures can be put with a high confidence level.

The AUC under the ROC curve is one of the most relevant parameters for assessing and rating the output efficiency of the classification and prediction models with a balanced sample condition; that is, the presence and the number of the absence of cases of HDs are roughly equal in the training and test data sets. However, if the data set contains unbalanced samples, the F-score is the most critical parameter for the consistency evaluation classification and prediction models. In our present study, the AUC would be an appropriate way to rate the classification and prediction models' output levels [67].

B. Kappa Statistics

Kappa error, or the interpretation of Cohen's Kappa Statistics, is used to determine the classifier's effectiveness. Output comparisons in classification algorithms can only yield incorrect results using the percentage of misses as a single predictor for precision [68]. Also, the focus must be given to the possibility of mistakes when making these decisions. In this sense, the Kappa error is a fair predictor for determining classifications, which may be due to chance. The Kappa error usually interprets values in between −1 and +1. If the classifiers measured Kappa value reaches '1', it is presumed the classifier's output is more realistic than by chance. Hence, Kappa error is a suggested parameter for calculation purposes in the performance analysis of classifiers. Thus, kappa statistics is a statistical measure of two data sets' 'inter-rate' conformity. The significant difference between 0 and 1 is representative of a better inter-rate agreement. For a chance of agreement, Kappa represents the convention's usual meaning. The agreement calculation is more rigorous than a simple percentage. Equation 10 describes the Kappa statistic.

$$\text{KAPPA(K)} = \frac{\text{Percentage\_of\_agreement}(P_A) - \text{Chance\_of\_agreement}(C_A)}{1 - \text{Chance\_of\_agreement}(C_A)} \quad (10)$$

Where K=1, implies the full inter-rate agreement between the classifier and ground truth, and K=0 is a certain probability of consensus. Comparisons of success in classification algorithms can yield misleading results only by using the percentage of misses for accuracy as the single meter. Even when making such calculations, the cost of error needs attention.

VIII. SIGNIFICANT FEATURES

TABLE III lists out the features, derived from the applications of four filter methods, in descending order of significance.

The results in TABLE III show that three tests namely thal (thalassemia), cp (chest pain) and ca (Number of major vessels colored by fluoroscopy), are the most significant ones to predict heart disease, excepting Relief-F which suggests 'restecg' as third important factor and cp as fifth-ranked. Information-Gain and Gain ratio show Oldpeak (ST Depression induced by exercise relative to rest) as the fourth important feature. The other two methods, namely Relief-F and One-R, suggest age and sex as the fourth important feature, respectively. Both Information-Gain and Gain ratio identifies thalach (Maximum Heart Rate Achieved) and exang as the fifth and sixth critical factors. Thus, these two methods have a similar ranking of features.

These methods show Chol (or cholesterol) as the most insignificant feature. This finding contradicts the views of doctors who opine show that measure of cholesterol and blood sugar in conjunction with age and ECG status as the significant features to detect heart disease. Almost all the feature selection methods show these features as the least important ones.

TABLE III. FEATURES IN DESCENDING ORDER OF SIGNIFICANCE

Information Gain Ranking Filter		**GainRatio AttributeEval**	
0.1887	thal	0.1893	thal
0.1726	cp	0.1732	cp
0.1592	ca	0.1638	ca
0.1242	oldpeak	0.1253	oldpeak
0.1006	thalach	0.1061	thalach
0.0846	exang	0.1015	exang

0.0803	age	0.081	age
0.0715	slope	0.0716	slope
0.0638	sex	0.0688	sex
0	chol	0	chol
0	fbs	0	fbs
0	trestbps	0	trestbps
0	restecg	0	restecg
ReliefF Attribute Eval M 1 – D 1 – K 10		**OneR Attribute Eval S 1 – F 10 – B 6**	
0.10565	thal	75.484	thal
0.09204	ca	74.194	ca
0.08032	restecg	74.194	cp
0.07419	sex	70.323	age
0.06645	cp	68.387	exang
0.05871	exang	66.452	thalach
0.03658	oldpeak	65.161	slope
0.02613	slope	61.29	sex
0.02387	age	59.355	oldpeak
0.02327	thalach	57.419	trestbps
0.01086	trestbps	55.484	fbs
0.0071	fbs	50.323	restecg
-0.00209	chol	49.677	chol

IX. HEART DISEASE (HD) PREDICTION

The MLT on the dataset considering all features (13) showed highest accuracy of 93.55% using Logistics Regression and Naïve Bayes. The precision level was 0.90, maximum sensitivity of 0.90, and specificity of 0.95. The authors obtained higher accuracy levels, i.e., when the features reduced from 13 to 10 by removing insignificant features obtained using Information Gain, Gain Ratio, Relief F and One R. However, the significance of bottom three features varied with the FS techniques. The authors applied MLT to all the subsets obtained from using different FS methods. The prediction accuracy increased to 96.77%, with sensitivity, specificity, Cohens Kappa and ROC-AUC of 0.90, 1, 0.92 and 0.95 respectively using ten important features identified by Information Gain and Gain Ratio methods and classified using Naïve Bayes. Similar exercise was carried out by reducing the features by one in every step and classifying them with the MLTs. The outcome is shown in Table 4. The result remained unchanged when chest pain (cp) was removed.

Thus, the best result in terms of accuracy, of 96.77%, was obtained using nine features obtained from Information Gain and Gain Ratio FS methods and classifying with NB technique. This showed that NB was the most preferred method with nine features.

The four most unimportant features showed by this FS method include cholesterol (Chol), chest pain (cp), blood pressure (Trestbps), and ECG (Restecg). The significant ones are – thal (thalassemia – defect type), ca (check for arterial blockages), oldpeak (ST Depression induced by exercise relative to rest), thalach (maximum heart rate achieved), exang (Exercise Induced Angina), age, slope (Slope of the peak exercise ST segment), sex, and fbs (Fasting Blood Sugar).

The authors carried out prediction using cholesterol (Chol), blood pressure (Trestbps), and ECG (Restecg) and found to be only 51.61%. This finding answers the research question that the most widely used approach of measuring cholesterol, blood pressure, and ECG does not lead to satisfactory prediction.

X. RESULTS AND DISCUSSION

The authors used an HD dataset to compare results obtained from the application of MLTs and a combination of various FS methods with MLTs for predicting disease risk. The feature selection methods included - Information Gain, Gain Ratio, Relief F and One R. Five classification algorithms for testing of the classification accuracy included state-of-art algorithms – non-ensemble methods namely, NB, Decision Tree (J48), SVM, LR, and ensemble techniques – RF, GDB, and ET.

The authors determined the validity of outcomes of the combination of FS and MLTs using performance metrics such as the Ratio of TP and TN (TP ÷ TN), where TP is True Positives and TN is True Negatives, Sensitivity or, Recall, Specificity, and precision measures namely, Cohen’s Kappa and AUC.

The results show 93.55 % as the highest accuracy of prediction with all features applying MLT. The three techniques, namely NB, SVM, and LR, predicted uniformly with all features. As the authors applied feature selection, it included the features in diminishing order of significance. The authors choose a subset based on the method of inclusion. In the process, they found that the exclusion of three features – cholesterol (Chol), blood pressure (Trestbps), and ECG (Restecg) improved the prediction level from 93.55% to 96.77% using Naïve Bayes technique.

Further, the authors found that dropping cp or chest-pain, as this was not relevant to asymptomatic patients, did not affect the accuracy levels. Thus, the authors answered the two research questions – whether the present widely used method of considering cholesterol (Chol), blood pressure (Trestbps), and ECG (Restecg) as significant features stands correct?, and what are the most important features to determine the heart disease (HD)?.

TABLE IV and TABLE V show performance metrics of classification algorithms with the number of features selected from the FS methods. According to these two tables, the highest precision value (precision=1) was obtained for the HD dataset with the top 9 features excluding cp selected by Information Gain and Gain Ratio and predicted with NB ML technique. The highest values of the different precision metrics include – sensitivity of 1, specificity of 0.95, ROC value of 0.93, and Kappa statistics value of 0.84 as shown in TABLE IV and TABLE V.

TABLE IV. ASSESSMENT OF FS METHODS FOR HD DATASET

Classifiers	**FS Method**	**Number of Features**	**Accuracy**	**Precision**	**ROC Area**	**Kappa Statistic**
NB		13 (all features)	93.55	0.90	0.93	0.85
	Information Gain	8	93.55	0.90	0.93	0.85
	Gain Ratio	8	93.55	0.90	0.93	0.85
	Relief F	8	93.55	1	0.90	0.84

	One R	8	93.55	0.90	0.93	0.85
	Information Gain	6	93.55	0.90	0.93	0.85
	Gain Ratio	6	93.55	0.90	0.93	0.85
	Relief F	6	90.32	0.89	0.88	0.77
	One R	6	90.32	0.89	0.88	0.77
	Information Gain	4	87.10	0.80	0.85	0.70
	Gain Ratio	4	87.10	0.80	0.85	0.70
	Relief F	4	83.87	0.78	0.80	0.62
	One R	4	80.65	0.80	85.24	0.71
LR		13	93.55	0.90	0.93	0.85
	Information Gain	8	93.55	1	0.90	0.84
	Gain Ratio	8	93.55	1	0.90	0.84
	Relief F	8	87.10	0.80	**0.85**	**0.70**
	One R	8	93.55	1	0.90	0.84
	Information Gain	6	90.32	0.82	0.90	0.78
	Gain Ratio	6	90.32	0.82	0.90	0.78
	Relief F	6	87.10	0.80	0.85	0.70
	One R	**6**	**93.55**	**1**	**0.90**	**0.84**
	Information Gain	4	87.10	0.80	0.85	0.70
	Gain Ratio	4	87.10	0.80	0.85	0.70
	Relief F	4	83.87	0.78	0.80	0.62
	One R	4	90.32	0.80	0.85	0.71
DT		13	77.42	0.62	0.78	0.52
	Information Gain	8	80.65	0.70	0.78	0.58
	Gain Ratio	8	80.65	0.70	0.78	0.56
	Relief F	8	58.06	0.41	0.61	0.19
	One R	8	80.65	0.70	0.78	0.56
	Information Gain	6	74.19	0.58	0.73	0.44
	Gain Ratio	6	74.19	0.58	0.73	0.44
	Relief F	6	87.10	0.80	0.85	0.70
	One R	6	83.87	0.73	0.83	0.64
	Information Gain	4	80.65	0.64	0.83	0.60
	Gain Ratio	4	80.65	0.64	0.83	0.60
	Relief F	4	90.32	1	0.85	0.76
	One R	4	80.65	0.75	0.75	0.53
SVM		13	93.55	1	0.90	0.84
	Information Gain	8	93.55	1	0.90	0.84
	Gain Ratio	8	93.55	1	0.90	0.84
	Relief F	8	90.32	0.89	0.88	0.77
	One R	8	93.55	1	0.90	0.84
	Information Gain	6	87.10	0.80	0.85	0.70
	Gain Ratio	6	87.10	0.80	0.85	0.70
	Relief F	6	90.32	0.89	0.88	0.77
	One R	6	90.32	0.89	0.88	0.77
	Information Gain	4	87.10	0.80	0.85	0.70
	Gain Ratio	4	87.10	0.80	0.85	0.70
	Relief F	4	87.10	0.88	0.83	0.69
	One R	4	87.10	0.80	0.85	0.71
RF		13	80.65	0.70	0.78	0.56
	Information Gain	8	74.19	0.58	0.73	0.44
	Gain Ratio	8	74.19	0.58	0.73	0.44
	Relief F	8	80.65	0.70	0.78	0.56
	One R	8	74.19	0.58	0.73	0.44
	Information Gain	6	87.10	0.75	0.88	0.72
	Gain Ratio	6	87.10	0.75	0.88	0.72
	Relief F	6	83.87	0.73	0.83	0.64
	One R	6	80.65	0.70	0.78	0.58
	Information Gain	4	77.42	0.62	0.78	0.52
	Gain Ratio	4	77.42	0.62	0.78	0.52
	Relief F	4	90.32	1	0.85	0.78
	One R	4	70.97	0.60	0.70	0.41
GDB		13	87.10	0.87	0.85	0.70
	Information Gain	8	77.42	0.80	0.78	0.52
	Gain Ratio	8	77.42	0.80	0.78	0.52
	Relief F	8	87.10	0.87	0.83	0.69
	One R	8	77.42	0.80	0.78	0.52
	Information Gain	6	77.42	0.80	0.78	0.52
	Gain Ratio	6	77.42	0.80	0.78	0.52
	Relief F	6	83.87	0.84	0.83	0.64
	One R	6	83.87	0.84	0.83	0.64
	Information Gain	4	80.65	0.82	0.80	0.58
	Gain Ratio	4	80.65	0.82	0.58	0.80
	Relief F	4	90.32	0.92	0.85	0.76
	One R	4	74.19	0.76	0.73	0.44
ET		13	90.32	0.90	0.88	0.77
	Information Gain	8	74.19	0.76	0.73	0.44
	Gain Ratio	8	77.42	0.80	0.78	0.52
	Relief F	8	67.74	0.72	0.68	0.33
	One R	8	7.42	0.80	0.78	0.52
	Information Gain	6	80.65	0.80	0.75	0.53
	Gain Ratio	6	80.65	0.80	0.75	0.53
	Relief F	6	87.10	0.87	0.85	0.70
	One R	6	83.87	0.84	0.83	0.64
	Information Gain	4	80.65	0.82	0.80	0.58
	Gain Ratio	4	80.65	0.82	0.80	0.58
	Relief F	4	90.32	0.92	0.85	0.76
	One R	4	77.42	0.78	0.75	0.50

The findings reveal that:

i. Ensemble methods did not perform better than other supervised machine learning techniques (MLT).

ii. Reduced features improved the prediction accuracy and showed better performance determined using Precision, Kappa and ROC/AUC.

iii. Feature selection (FS) methods varied in suggesting the order of feature significance.

iv. Integration of FS with ensemble and other supervised MLT leads to better diagnosis that limiting the analysis to any particular technique.

v. Information Gain and Gain Ratio were found to superior than Relief-F and One-R FS methods.

TABLE V. COMPARISON OF SENSITIVITY AND SPECIFICITY

Algorithm	Feature Selection Method	No of Features	Sensitivity/ Recall	Specificity
NB		13 (all factors)	0.90	0.95
	Information Gain	8	0.90	0.95
	Gain Ratio	8	0.90	0.95
	Relief F	8	1	0.91
	One R	8	0.90	0.95
	Information Gain	6	0.90	0.95
	Gain Ratio	6	0.90	0.95

	Relief F	6	0.89	0.91
	One R	6	0.89	0.91
	Information Gain	4	0.80	0.90
	Gain Ratio	4	0.80	0.90
	Relief F	4	0.78	0.86
	One R	4	0.70	0.86
LR		13	0.90	0.95
	Information Gain	8	1	0.91
	Gain Ratio	8	1	0.91
	Relief F	8	0.80	0.90
	One R	8	1	0.91
	Information Gain	6	1	0.91
	Gain Ratio	6	1	0.91
	Relief F	6	0.80	0.91
	One R	6	1	0.91
	Information Gain	4	0.80	0.90
	Gain Ratio	4	0.80	0.90
	Relief F	4	0.78	0.86
	One R	4	0.89	0.91
DT		13	0.62	0.89
	Information Gain	8	0.70	0.86
	Gain Ratio	8	0.70	0.86
	Relief F	8	0.41	0.79
	One R	8	0.70	0.86
	Information Gain	6	0.70	0.86
	Gain Ratio	6	0.70	0.86
	Relief F	6	0.80	0.91
	One R	6	0.73	0.90
	Information Gain	4	0.64	0.94
	Gain Ratio	4	0.64	0.94
	Relief F	4	1	0.88
	One R	4	0.80	0.91
SVM		13	1	0.91
	Information Gain	8	1	0.91
	Gain Ratio	8	1	0.91
	Relief F	8	0.89	0.91
	One R	8	1	0.91
	Information Gain	6	1	0.91
	Gain Ratio	6	1	0.91
	Relief F	6	0.89	0.91
	One R	6	0.89	0.91
	Information Gain	4	0.80	0.90
	Gain Ratio	4	0.80	0.90
	Relief F	4	0.88	0.87
	One R	4	0.67	0.89
RF		13	0.70	0.86
	Information Gain	8	0.58	0.84
	Gain Ratio	8	0.58	0.84
	Relief F	8	0.70	0.86
	One R	8	0.58	0.84
	Information Gain	6	0.58	0.84
	Gain Ratio	6	0.58	0.84
	Relief F	6	0.73	0.90
	One R	6	0.70	0.86
	Information Gain	4	0.62	0.89
	Gain Ratio	4	0.62	0.89
	Relief F	4	1	0.88
	One R	4	0.55	0.80
GDB		13	0.80	0.91
	Information Gain	8	0.62	0.89
	Gain Ratio	8	0.62	0.89
	Relief F	8	0.88	0.87
	One R	8	0.62	0.89
	Information Gain	6	0.62	0.89
	Gain Ratio	6	0.62	0.89
	Relief F	6	0.73	0.90
	One R	6	0.73	0.90
	Information Gain	4	0.67	0.89
	Gain Ratio	4	0.67	0.89
	Relief F	4	1	0.88
	One R	4	0.58	0.84
ET		13	0.89	0.91
	Information Gain	8	0.58	0.84
	Gain Ratio	8	0.58	0.84
	Relief F	8	0.50	0.82
	One R	8	0.62	0.89
	Information Gain	6	0.62	0.89
	Gain Ratio	6	0.62	0.89
	Relief F	6	0.80	0.91
	One R	6	0.73	0.90
	Information Gain	4	0.67	0.89
	Gain Ratio	4	0.67	0.89
	Relief F	4	1	0.88
	One R	4	0.64	0.85

XI. CONCLUSION

Machine learning algorithms can predict the outcome with different levels of accuracy. Studies reveal that prediction accuracy is affected by type and size of dataset, nature of classifying technique (supervised and un-supervised) and the presence of unwanted features in the dataset. Thus, feature selection (FS) is an essential step in processing data in data mining studies and no single technique can be preferred over another. Since errors in medical treatment have serious implications, as a necessary condition, it is crucial to integrate FS methods and MLTs. The FS should precede classifications for diagnosis of HD.

This study addresses the myth – HD can be predicted well with data on the level of cholesterol, blood pressure, and ECG. The results show that the highest prediction accuracy using most MLTs is only to the tune of 56.61% with these three features.

In the patients' treatment, type 1 error (False positive), and type 2 error (False negative) are equally important. A person not diagnosed with the disease will cause fatality, while the treatment of a person without the disease will lead to negative results, and as such, the performance metrics need evaluation by a researcher. In this study, the authors obtained a sensitivity of 0.90, specificity of 1, ROC value of 0.95, and Kappa statistics value of 0.92. These are acceptable, and the results concluded stand validated.

In general, previous studies showed that Random-Forest had delivered higher accuracy levels compared to other widely used machine learning techniques (MLTs), and Logistics Regression showed consistency in prediction results. However, no method could claim the overall superiority over others. In this study, the authors found that accuracy levels improved with fewer features, and Naïve Bayes surpassed the accuracy levels of the other techniques. Thus, the authors conclude:

1. Feature selection is a necessary condition – to decide whether to keep or discard features, and
2. Different machine learning techniques (MLTs) can predict with higher accuracy given specific pre-requisites, such as the dataset's size, number of features, and records per category, and hence simultaneous combination of FS and MLT are proposed.

Thus, the authors propose an extrinsic ensemble of feature selection and machine learning techniques to predict diseases, leading to a reduction in errors. However, the attempt to optimize the prediction accuracy by fixing the constraints regarding specificity and sensitivity through a simulation of

FS and MLT would yield a better result. The authors propose such an attempt as future work.

REFERENCES

[1] S. Itani, F. Lecron, and P. Fortemps, “Specifics of medical data mining for diagnosis aid: A survey,” Expert. Syst. Appl., vol. 118, pp. 300-314, 2019.

[2] D. Tomar and S. Agarwal, “Feature selection based least square twin support vector machine for diagnosis of heart disease,” Int. J. Biosci. Biotechnol., vol. 6, pp. 69-82, 2014.

[3] M. Durairaj and V. Ranjani, “Data mining applications in healthcare sector: a study,” Int. J. Sci. Technol. Res., vol. 2, pp. 29-35, 2013.

[4] O. Sagi and L. Rokach, “Ensemble learning: A survey,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 8, 2018.

[5] I. Partalas, G. Tsoumakas, and I.P. Vlahavas, “Focused Ensemble Selection: A Diversity-Based Method for Greedy Ensemble Selection,” in ECAI, August 2008, pp. 117-121.

[6] O. Araque, I. Corcuera-Platas, J.F. Sánchez-Rada, and C.A. Iglesias, “Enhancing deep learning sentiment analysis with ensemble techniques in social applications,” Expert. Syst. Appl., vol. 77, pp. 236-246, 2017.

[7] K. Shankar, S.K. Lakshmanaprabu, D. Gupta, A. Maseleno, and V.H.C. De Albuquerque, “Optimal feature-based multi-kernel SVM approach for thyroid disease classification,” J. Supercomput., vol. 76, pp. 1128-1143, 2020.

[8] Z. Zhang, J. Dong, X. Luo, K.S. Choi, and X. Wu, “Heartbeat classification using disease-specific feature selection,” Comput. Biol. Med., vol. 46, pp. 79-89, 2014.

[9] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, “A review of feature selection methods on synthetic data,” Knowl. Inform. Syst., vol. 34, pp. 483-519, 2013.

[10] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, Feature selection for high-dimensional data. Cham: Springer International Publishing, 2015.

[11] Y. Saeys, S. Degroeve, D. Aeyels, Y. Van de Peer, and P. Rouzé, “Fast feature selection using a simple estimation of distribution algorithm: a case study on splice site prediction,” Bioinformatics, vol. 19, pp. ii179-ii188, 2003.

[12] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, “Feature selection: An ever evolving frontier in data mining,” in Feature selection in data mining, PMLR, May 2010, pp. 4-13.

[13] Y. Saeys, I. Inza, and P. Larranaga, “A review of feature selection techniques in bioinformatics,” Bioinformatics, vol. 23, pp. 2507-2517, 2007.

[14] K. Rajeswari, V. Vaithiyanathan, and T.R. Neelakantan, “Feature selection in ischemic heart disease identification using feed forward neural networks,” Procedia Eng., vol. 41, pp. 1818-1823, 2012.

[15] Y.J. Park, S.H. Chun, and B.C. Kim, “Cost-sensitive case-based reasoning using a genetic algorithm: Application to medical diagnosis,” Artif. Intell. Med., vol. 51, pp. 133-145, 2011.

[16] R. Varatharajan, G. Manogaran, and M.K. Priyan, “A big data classification approach using LDA with an enhanced SVM method for ECG signals in cloud computing,” Multimed. Tool. Appl., vol. 77, pp. 10195-10215, 2018.

[17] İ. Babaoglu, O. Findik, and E. Ülker, “A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine,” Expert. Syst. Appl., vol. 37, pp. 3177-3183, 2010.

[18] C.M. Florkowski, “Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests,” Clin. Biochem. Rev., vol. 29, pp. S83, 2008.

[19] A. Hazra and N. Gogtay, “Biostatistics series module 7: the statistics of diagnostic tests,” Indian J. Dermatol., vol. 62, pp. 18, 2017.

[20] S. Bashir, Z.S. Khan, F.H. Khan, A. Anjum, and K. Bashir, “Improving heart disease prediction using feature selection approaches,” in 2019 16th international bhurban conference on applied sciences and technology, IEEE, January 2019, pp. 619-623.

[21] A.K. Verma, S. Pal, and S. Kumar, “Comparison of skin disease prediction by feature selection using ensemble data mining techniques,” Inform. Med. Unlocked, vol. 16, 2019.

[22] R. Duangsoithong and T. Windeatt, “Relevant and redundant feature analysis with ensemble classification,” in 2009 Seventh International Conference on Advances in Pattern Recognition, IEEE, February 2009, pp. 247-250.

[23] M.C. Tu, D. Shin, and D. Shin, “Effective diagnosis of heart disease through bagging approach,” in 2009 2nd international conference on biomedical engineering and informatics, IEEE, October 2009, pp. 1-4.

[24] K. Srinivas, B.K. Rani, and A. Govrdhan, “Applications of data mining techniques in healthcare and prediction of heart attacks,” Int. J. Comput. Sci. Eng., vol. 2, pp. 250-255, 2010.

[25] P.K. Anooj, “Clinical decision support system: Risk level prediction of HD using weighted fuzzy rules,” J. King Saud Univ. - Comput. Inf. Sci., vol. 24, pp. 27-40, 2012.

[26] E.M. Karabulut and T. İbrikçi, “Effective diagnosis of coronary artery disease using the rotation forest ensemble method,” J. Med. Syst., vol. 36, pp. 3011-3018, 2012.

[27] M. Shouman, T. Turner, and R. Stocker, “Applying k-nearest neighbour in diagnosing heart disease patients,” Int. J. Inf. Educ. Technol., vol. 2, pp. 220-223, 2012.

[28] M. Shouman, T. Turner, and R. Stocker, “Integrating decision tree and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients,” in Proceedings of the International Conference on Data Science, 2012, pp. 1.

[29] V. Chaurasia and S. Pal, “Early prediction of heart diseases using data mining techniques,” Caribb. J. Sci. Technol., vol. 1, pp. 208-217, 2013.

[30] S. Hari Ganesh and M. Gajenthiran, “Comparative study of data mining approaches for prediction heart diseases,” IOSR J. Comput. Eng., vol. 4, pp. 36-39, 2014.

[31] S. Bashir, U. Qamar, and M.Y. Javed, “An ensemble based decision support framework for intelligent heart disease diagnosis,” in International conference on information society, IEEE, November 2014, pp. 259-264.

[32] J. Patel, D. TejalUpadhyay, and S. Patel, “Heart disease prediction using machine learning and data mining technique,” Heart. Dis., vol. 7, pp. 129-137, 2015.

[33] O.W. Samuel, G.M. Asogbon, A.K. Sangaiah, P. Fang, and G. Li, “An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction,” Expert. Syst. Appl., vol. 68, pp. 163-172, 2017.

[34] H. Kahramanli and N. Allahverdi, “Design of a hybrid system for the diabetes and heart diseases,” Expert. Syst. Appl., vol. 35, pp. 82-89, 2008.

[35] M.K. Priyan and G.U. Devi, “Energy efficient node selection algorithm based on node performance index and random waypoint mobility model in internet of vehicles,” Cluster Comput., vol. 21, pp. 213-227, 2018.

[36] N.M. Nawi, R. Ghazali, and M.N.M. Salleh, “The development of improved back-propagation neural networks algorithm for predicting patients with heart disease,” in International conference on information computing and applications, Heidelberg, Berlin: Springer, October 2010, pp. 317-324.

[37] S. Paredes, T. Rocha, P. De Carvalho, J. Henriques, M. Harris, and J. Morais, “Long term cardiovascular risk models’ combination,” Comput. Meth. Programs. Biomed., vol. 101, pp. 231-242, 2011.

[38] A.K. Chaudhuri, D. Sinha, and K.S. Thyagaraj, “Review of efficiency of k-means algorithm on studies related to cardio vascular diseases”, 2015.

[39] R.C. Pasternak, S.M. Grundy, D. Levy, and P.D. Thompson, “Task Force 3. Spectrum of risk factors for coronary heart disease,” J. Am. Coll. Cardiol., vol. 27, pp. 978-990, 1996.

[40] S.M. Grundy, “Primary prevention of coronary heart disease: integrating risk assessment with intervention,” Circulation, vol. 100, pp. 988-998, 1999.

[41] P. Greenland, M.D. Knoll, J. Stamler, J.D. Neaton, A.R. Dyer, D.B. Garside, and P.W. Wilson, “Major risk factors as antecedents of fatal and nonfatal coronary heart disease events,” Jama, vol. 290, pp. 891-897, 2003.

[42] S. Prakash, K. Sangeetha, and N. Ramkumar, “An optimal criterion feature selection method for prediction and effective analysis of heart disease,” Cluster Comput., vol. 22, pp. 11957-11963, 2019.

[43] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, “A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease,” in 2017 IEEE Symposium on Computers and Communications, IEEE, July 2017, pp. 204-207.

[44] N.C. Long, P. Meesad, and H. Unger, “A highly accurate firefly based algorithm for heart disease prediction,” Expert. Syst. Appl., vol. 42, pp. 8221-8231, 2015.

[45] J. Nahar, T. Imam, K.S. Tickle, and Y.P.P. Chen, “Association rule mining to detect factors which contribute to heart disease in males and females,” Expert. Syst. Appl., vol. 40, pp. 1086-1093, 2013.

[46] I.H. Witten and E. Frank, “Data mining: practical machine learning tools and techniques with Java implementations,” ACM SIGMOD Rec., vol. 31, pp. 76-77, 2002.

[47] I. Kononenko, “Estimating attributes: Analysis and extensions of RELIEF,” in European conference on machine learning, Heidelberg, Berlin: Springer, April 1994, pp. 171-182.

[48] K. Kira and L.A. Rendell, “A practical approach to feature selection,” in Machine learning proceedings, Morgan Kaufmann, 1992, pp. 249-256.

[49] P. Yildirim, “Filter based feature selection methods for prediction of risks in hepatitis disease,” Int. J. Mach. Learn. Comput., vol. 5, pp. 258, 2015.

[50] A.G. Karegowda, A.S. Manjunath, and M.A. Jayaram, “Comparative study of attribute selection using gain ratio and correlation based feature selection,” Int. J. Inf. Technol. Knowl. Manag., vol. 2, pp. 271-277, 2010.

[51] T.R. Patil and S.S. Sherekar, “Performance Analysis of J48 and J48 Classification Algorithm for Data Classification,” Int. J. Comput. Sci. Appl., vol. 6, 2013.

[52] G. Dimitoglou, J.A. Adams, and C.M. Jim, “Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability,” 2012, arXiv preprint arXiv:1206.1121.

[53] J.R. Quinlan, C4.5: programs for machine learning. Elsevier, 2014.

[54] C. Cortes and V. Vapnik, “Support-vector networks,” Mach. Learn., vol. 20, pp. 273-297, 1995.

[55] M. Mertik, P. Kokol, and B. Zalar, “Gaining features in medicine using various data-mining techniques,” in IEEE 3rd International Conference on Computational Cybernetics, IEEE, April 2005, pp. 21-24.

[56] N. Barakat, A.P. Bradley, and M.N.H. Barakat, “Intelligible support vector machines for diagnosis of diabetes mellitus,” IEEE Trans. Inform. Tech. Biomed., vol. 14, pp. 1114-1120, 2010.

[57] G. Suganya and D. Dhivya, “Extracting diagnostic rules from support vector machine,” J. Comput. Appl., vol. 4, 2011.

[58] S. Balakrishnan, R. Narayanaswamy, N. Savarimuthu, and R. Samikannu, “SVM ranking with backward search for feature selection in type II diabetes databases,” in 2008 IEEE International Conference on Systems, Man and Cybernetics, IEEE, October 2008, pp. 2628-2633.

[59] A.R. Cannon, G.W. Cobb, B.A. Hartlaub, J.M. Legler, R.H. Lock, T.L. Moore, and J. Witmer, STAT2: building models for a world of data. W.H. Freeman, 2013.

[60] L. Breiman, “Random forests,” Mach. Learn., vol. 45, pp. 5-32, 2001.

[61] R.J. Marshall, “The use of classification and regression trees in clinical epidemiology,” J. Clin. Epidemiol., vol. 54, pp. 603-609, 2001.

[62] N. Landwehr, M. Hall, and E. Frank, “Logistic model trees,” Mach. Learn., vol. 59, pp. 161-205, 2005.

[63] A. Sharaff and H. Gupta, “Extra-tree classifier with metaheuristics approach for email classification,” in Advances in Computer Communication and Computational Sciences, Singapore: Springer, 2019, pp. 189-197.

[64] R. Kannan and V. Vasanthi, “Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease,” in Soft Computing and Medical Bioinformatics, Singapore: Springer, 2019, pp. 63-72.

[65] S. Chen, J. Xu, L. Chen, X. Zhang, L. Zhang, and J. Li, “A Regularization-Based eXtreme Gradient Boosting Approach in Foodborne Disease Trend Forecasting,” Stud. Health. Technol. Inform., vol. 264, pp. 930-934, 2019.

[66] A. Ray and A.K. Chaudhuri, “Smart healthcare disease diagnosis and patient management: Innovation, improvement and skill development,” Mach. Learn. Appl., vol. 3, 2021.

[67] L.F. Chalak, L. Pavageau, B. Huet, and L. Hynan, “Statistical rigor and kappa considerations: which, when and clinical context matters,” Pediatr. Res., vol. 88, pp. 5, 2020.

[68] A. Ozcift, “SVM feature selection based rotation forest ensemble classifiers to improve computer-aided diagnosis of Parkinson disease,” J Med Syst, vol. 36, pp. 2141-2147, 2012.

[69] K.M. Anderson, P.M. Odell, P.W. Wilson, and W.B. Kannel, “Cardiovascular disease risk profiles,” Am. Heart J., vol. 121, pp. 293-298, 1991.