International Journal of Cognitive Informatics and Natural Intelligence
Volume 15 • Issue 4 • October-December 2021

A Comprehensive Performance Analysis of Various Classifier Models for Coronary Artery Disease Prediction

Baranidharan Balakrishnan, SRM Institute of Science and Technology, India
Vinoth Kumar C. N. S., SRM Institute of Science and Technology, India

ABSTRACT

Cardio vascular diseases (CVD) are the major reason for the death of the majority of the people in the world. Earlier diagnosis of disease will reduce the mortality rate. Machine learning (ML) algorithms are giving promising results in the disease diagnosis, and they are now widely accepted by medical experts as their clinical decision support system. In this work, the most popular ML models are investigated and compared with one other for heart disease prediction based on various metrics. The base classifiers such as support vector machine (SVM), logistic regression, naïve Bayes, decision tree, k-nearest neighbour are used for predicting heart disease. In this paper, bagging and boosting techniques are applied over these individual classifiers to improve the performance of the system. With the Cleveland and Statlog datasets, naive Bayes as the individual classifier gives the maximum accuracy of 85.13%and 84.81%, respectively. Bagging technique improves the accuracy of the decision tree, which is identified as a weak classifier by 7%, and it is a significant improvement in identifying CVD.

KEYWORDS

Classifiers, Data Mining Techniques, Heart Disease Prediction, Machine Learning Techniques

I. INTRODUCTION

In recent years, a large volume of medical data is being generated in the hospital and health care institutions due to the extensive use of digital technologies. Big data analytics methods will extract a lot of useful information from this voluminous data. Javad Hassannataj Joloudari et al (2020) analyzed that Data science has significant growth by taking into the reach of big data for smart diagnose, disease avoidance, and policy-making in the medical sector. Raghupathi et al (2010) experimented predictive models built on this data will help the clinic in early diagnosis of disease, reduce cost, and improve treatment and overall clinical experience of the patients.

Cardio Vascular Disease (CVD) is the collective term to represent any form of heart-related diseases Ahmad, G, Wang et al (2019) (2019). It includes high blood pressure, coronary artery disease, peripheral artery disease, cerebrovascular disease, etc… Coronary Artery Disease (CAD) is the state of arteries carrying blood to the heart muscle is narrowing down due to plaque built in it. CAD is said to be an important killer disease in the entire universe by the World Health Organization (WHO). From a survey of the 2015 article, it is mentioned that about 110 million peoples were infected with CAD. It confronts that 17.9 million deaths, out of 31% deaths occurred in 2016 (World Health

DOI: 10.4018/IJCINI.20211001.oa36

Organization, 2017). The early conclusion of CAD hazard will rapidly increase the recommended treatment protocol and enormously enlarges the recovery speed of the patients.

Mostly, the heart related diseases are identified through Electrocardiogram (ECG) tests. Any irregularities in the heart can be identified using ECG by medical experts Acharya U et al (2014) easily. But in some rare cases, the ECG also doesn't track the exact brutality of the CAD. Another popular way of identifying heart disease is by using Angiogram. But angiogram is the invasive method and economically costlier too. The high cost of Angiogram makes it less affordable for the economically weaker section of the people. To make the diagnosis system widely applicable and economically affordable a new less complex, minimal effort and exact diagnosis model should be built with the assistance of ongoing technological advancement.

AI [ML] based prescient frameworks are being created by Tech organizations (Indo-Asian News Service, 2018; Vincent, 2018) and academic institutions along with their accomplice emergency clinics. The most popular classification techniques used are Naive Bayes, Decision Tree, Simple Logistic Regression, Support Vector Machine, Artificial Neural Networks (ANNs). The increased numbers of cataloging models were created in the form of CAD diagnosis utilizing the previously mentioned systems. Be that as it may, the vast majority are newly created data sets from UCI storehouse. Coronary illness UCI data sets Andras Janosi et al (2015) contains 14 factors where 13 are free factors and 1 dependent factor.

(i) Logistic Regression: Logistic regression (LR) is the most straightforward of the considerable number of classifiers and computes the probability value between 0 and 1 for the given input. If the probability value is 0.5 or more then it is classified as class 1 otherwise it classifies the input to another class 0. The sigmoid function is used to compute the probability value between 0 or 1. LR utilizes logic or additionally called score based on a probabilistic strategy for distinguishing the class of new input.

$$p\left(z\right) = \frac{1}{1 + e^{-z}} \tag{1}$$

Equation (1) depicts the sigmoid function used for computing the probability value between 0 and 1.z represents given input to the sigmoid function. If the result of the sigmoid function is within 0.5, the given input will be assigned class 0 and if the probability output is between 0.5 and 1, class 1 is assigned.

(ii) Naïve Bayes: Naïve-Bayes classifier is based on the Bayesian theorem. Equation (2) depicts the Bayesian model where the event X can be predicted given the occurrence of the event Y. Thus Bayesian theorem is the conditional probability theorem. Naïve Bayes model is based on the Bayesian concept and it is also based on conditional probability. Equation (3) depicts the Naive Bayes model where the term x1, x2, etc denotes each given input features and n represents the total number of features. Naive Bayes differs from the Bayesian theorem in two aspects: (i) Since the denominator is the same for all the classes, the denominator term is removed (ii) Assignment or equal to sign is replaced with a directionally proportional symbol. It is based on the supposition that each attribute or highlight is autonomous of each other and has its effect on the yielded output. On account of its restrictive independence of features, it is reasonable for high dimensional models. Although it is a basic classifier for now and then its performance is far better than advanced classifiers.

$$P\left(\frac{X}{Y}\right) = \frac{P\left(\frac{Y}{X}\right)P\left(X\right)}{P\left(Y\right)} \tag{2}$$

$$P\left(\frac{y}{x_1, x_2, \ldots x_n}\right) \alpha \; P(y) \prod_{i=1}^{n} P\left(\frac{x_i}{y}\right) \tag{3}$$

(iii) Decision Tree: Decision Tree creates the exact needed model with the relevant needed tree structure. In a Similar traditional tree structure, it equipped with root hub, intermediate, and leaf hubs. The root hub symbolizes the base component informational index division; other significant features are situated in the following levels of the tree structure. DTs continuously divide the data into subsets until the subset can be identified or termed by an already created label. DTs are used successfully used in both classification and regression problems. It is also called a Classification and Regression tree (CART).

(iv) K Nearest Neighbour: K Nearest Neighbour (KNN) is a non-parametric classifier calculation. It distinguishes a novel data based on its calculated space with the K nearest previously classified information. At this point when the data set is gigantic, KNN provides preferred execution and routine over the compared generality of other different classifiers.

(v) Support Vector Machine: Support Vector Machine is the huge edge classifier that characterizes the +ve and -ve information focuses on a bigger limit between them. SVM is solid classifiers which don't experience the ill effects of overfitting issue not at all like other comparative classifiers.SVM combines theoretical machine learning, kernel concept from mathematics, and optimization concept in the right way to minimize loss. It maximizes the boundary space between different classes through support vectors which are again from the training data set.

(vi)Ensemble model: Ensemble model Dietterich et al (2000) is the collection of one or more classifiers. Three methods of creating an ensemble model: (i) Bagging, is said to be the homogeneous or same sort of classifiers are utilized and an ultimate choice depends on the vote from every singular classifier. In Bagging all the classifiers are independent of each other and executed parallelly. The final decision is based on the voting procedure. Figure 1 depicts bagging algorithm.

(ii) Boosting is likewise an ensemble model similar way of bagging yet classifiers are requested in the arrangement. The presentation of the past model influences the upcoming model. Also, it gives importance to the weight parameter. Figure 2 outlines boosting algorithm. (iii) Voting Classifier, where it gathers heterogeneous or different types of classifiers. For example, different classifiers like SVM, Naïve Bayes, Logistic regression can be combined. The final output depends on the voting procedure.

II. LITERATURE SURVEY

Javad Hassannataj Joloudari et al (2020) discovered Database knowledge Discovery (KDD) is an optimal model to find the exact yield on disease diagnosis from the present healthcare scenario. The main challenge in the above said optimal model is the feature selection procedure, thus it helps to choose the finest subset of the datasets which it got trained earlier. One approach to precisely analyze this infection is to utilize data mining strategies to construct a proper and vigorous model that is more dependable than clinical imaging apparatuses, remembering angiography for the field of determination of coronary illness (Alizadehsani et al., 2019; Amin et al., 2019; Zipes et al., 2018). Through these strategies, the determination of the subset of features as indicated by their request priorities. For this reason, the subset of features is positioned from the least essential to the most significant because of the various weightings to the features related to the association models that these features were allocated to the yielded test simulator. At long last, among the classification models utilized in this study, acquiring the most suitable subset highlight by arbitrary trees model

Figure 1. Bagging Algorithm

Bagging Algorithm:

```
Dataset D= {d1, d2, d3, .... dn}
Set of classifier C = {c1, c2, c3, ... cn}
Ensemble Classifier list EC = { }

X = Training samples, where X is a subset of D
Y = Testing samples where Y is also a subset of D
R = number of training samples
S = number of testing samples
for i=1 to R
        T(i) = Bootstrap a sample randomly with replacement from the set X
        Train the classifier C(i) using T(i)
        Add the trained classifier C(i) to set EC
end
for i=1 to S
        for j = 1 to length of E
                Temp(j) = Classify S(i) using EC(j)
                Output = max(Temp(j))
        end
end
```

with the best characterization set and the most precise arrangement of coronary illness determination is the primary reason for this investigation.

Gulzar Ahmad et al (2019) analyzed various information examination methods, and some of them depend on AI, statistics, data abstraction, decision support system, and master framework N. Cheung (2001). Master framework methods have been utilized in the most recent couple of years in clinical investigation. They increment the diagnosis exactness and reducing the expenses M. Neshat et al (2009). At present, AI is being utilized to analyze various types of clinical issues. Intellectual frameworks are being created to determine the medicals issues A. Sardesai et al (2014). The fuzzy inference system (FIS) is a powerful master framework to investigate the issues and give their solutions. FIS is helpful where odds of vulnerability may arouse. It is utilized in each field of life, for example, programmed mechanical technology, businesses, PC sciences, clinical frameworks, climate determining, agribusiness, etc.

Alizadehsani et al (2019) found that the top reason for death in the world is coronary artery infection. Early recognition of CAD is basic to maintain a strategic distance from the further increment in the risk. Coronary angiography is required to decisively analyze CAD. In any case, it is obtrusive and may prompt different intricacies, for example, artery dissection, arrhythmia, and even demise. Besides, picture-based detection techniques are not pertinent for screening an enormous population. Because of these inadequacies and the hazardous nature of angiography, researchers have been persistently searching for non-invasive, prudent, quick, and solid procedures for the early location of CAD. ML calculations are a portion of the procedures utilized for this reason. ML-based methods have been effectively applied to different kinds of CAD datasets. These calculations have exhibited promising execution in the discovery and treatment of CAD. This investigation exhaustively audits how, when, and where ML strategies have been applied for CAD detection. In this way, giving the best outcomes and techniques can be useful for these people. We endeavor to discover answers to (i) the best performing ML strategy and (ii) locate the best sort of CAD data which will yield the

Figure 2. Boosting Algorithm

Boosting Algorithm:

```
Dataset D={d1,d2,d3, ... dn}
Set of Classifiers, C = {c1, c2, c3, ...cn}
Ensemble Classifier list, EC = { }
X = Training sample, where X is a subset of D
Y = Testing sample, where Y is also a subset of D
R = number of training samples
S = number of testing samples

SS = Select a random subset from X
M = { }
for i =1 to R
        if i>1
        s(i) = Set of incorrectly classified instances of M(i-1) + SS
        M(i) = Model trained using C(i) on S(i)
        Add M(i) to EC
end
for j = 1 to S
        Temp(j) = Y classified by E(i)
        Output = max(Temp(j): j=1,2, ..., n)

end
```

most noteworthy order results. By knowing the responses to these inquiries, specialists, doctors, governments, and patients can make different decisions about utilizing appropriate ML techniques.

Wang et al (2019) have developed an improved classifier based on a genetic algorithm in a recurrent fuzzy neural network. Though the system they developed is a complex one, the accuracy of around 97.78% they have achieved.

Alizadehsani et al (2013) had built the different sets of combination models for CAD reliant on Sequential Minimal Optimization (SMO), Bagging, Artificial Neural Networks, and Naive Bayes. Thus it grouped up with the SMO and Bagging where it is providing the more accurate degree of eighty-nine percentages and the Artificial neural network at eighty-five percentages, where Naive Bayes at more low.

Srinivasan et al (2010) had created a heart care system where it embeds 15 properties for contrasting the grimness of individual members where they working in coal mine shafts in Singaneri, Andhra Pradesh. Distinguished and various set of classifiers from the decision tree provides the best results comparatively.

Heart rate variability (HRV) is said to be the primary criterion for the cause of CAD by Melillo et al (2015) it is confined that the HRV projecting model finds out the severity of cardiovascular disease in a superior manner than even Echo graphic parameters do.

An Intelligent Heart Disease Prediction System was created by Palaniappan et al (2008) dependent on Naïve Bayes, Neural Network, and Decision Tree. It is inferred that Naïve Bayes has given higher exactness followed by Neural Network and Decision Tree. The investigation was directed more than 909 examples from UCI heart disease repository with an equivalent split of preparation and testing the same set.

Pouriyeh et al (2017) have done the comparative research metrics on the various classifiers over the Cleveland data set. Naive Bayes, Support Vector Machine, Radial Basis Function, Multi-Layer Perceptron, K Nearest Neighbour, Single Conjunctive Rule Learner, and Decision Tree Other than the ensemble techniques like bagging, boosting and stacking; it was applied over many individual classifiers to improve the outcomes. At last, they presumed that SVM upgraded with boosting strategy gives the most elevated exactness than the remaining.

Uyar et al (2017) have developed an improved classifier based on a genetic algorithm in a recurrent fuzzy neural network. Though the system they developed is a complex one, the accuracy of around 97.78% they have achieved.

The individual classifier algorithms give better results for particular data sets and fail to achieve the same for other data sets. So, here in these papers two major datasets: Cleveland and Statlog are used for comparing the classifier models. In the above literature survey, it is observed that there is no work done on exploring the power of individual classifier using bagging and boosting, and the effect of these techniques on weak classifiers are not at all considered. This led to a major research gap in CVD prediction. So, this research investigation was carried based on two main objectives,

1. Identifying the best individual classifier and the effect of ensemble techniques like bagging and boosting over it. Most of the literature survey have implemented and tested their algorithms only on Cleveland dataset. In this paper, apart from Cleveland, Statlog dataset is also used for testing and identifying the best individual classifier.
2. The study of the effect of bagging and boosting weak classifiers is an important one for designing a classifier system for the biomedical systems. This paper has made a detailed analysis of the improvement of the weak classifier using bagging and boosting techniques.

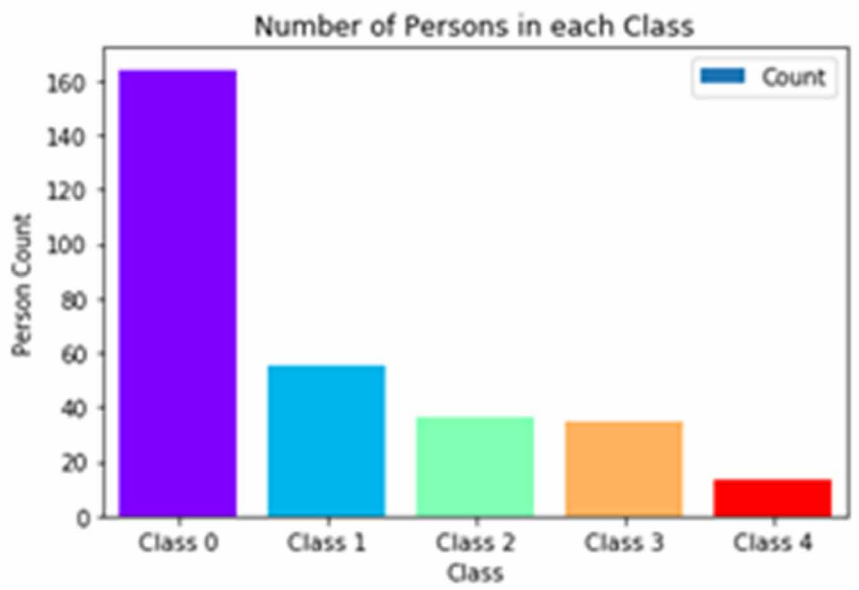
III. MEDICAL DATASETS

The vast majority of the AI specialists utilize the UCI coronary illness data set (Blake & Merz, 2015) which is made from four distinct sources: (i) Cleveland, (ii) Hungarian, (iii) Zurich, and (iv) Basel. Primarily, ML researchers use Cleveland data set collection because of its refinement. Originally, Cleveland dataset has 76variables for 303 records. But, out of 76 variables, 14 variables are identified more relevant to the coronary illness, they are Age, Gender, Chest pain, Resting pulse (Trestbps), Cholesterol, Fasting glucose (FBS), Restecg, Thalach, Exang, Old peak, slope, Ca, Thal and Class. Class is the yielded output variable which takes the qualities from 0 to 4 where 0 represents no coronary illness and the values 1 to 4 represent the seriousness of coronary illness in the increasing order. Figure 3(a) shows the number of records in every Class of Cleveland data set.

So it can be observed that out of 303 records, 164 people are not having any heart illness and the remaining 139 are having some form of heart disease. For our experimental procedure, we are having only two classes. Class 0 has no heart diseases and Class 1 with heart diseases.

In this research along with Cleveland data, Statlog dataset is also used for validating the classifiers. Statlog dataset is also referred to as the pure dataset since there are no missing values in it. From Figure 3(b), it is observed that there is a total of 270 patients record in Statlog dataset. Of them, 150 are not having any heart ailment and the remaining 120 are having some form of a heart ailment.

Figure 3. Number of Persons in each class of Cleveland and Statlog Dataset



IV. EXPERIMENTAL SETUP

The classifiers utilized for correlation in this research are conveyed and tried in the framework with the composition of the Intel i5 processor seventh era, 8GB RAM, Windows 10 working OS, and in Jupyter notebook environment. The inbuilt classifier models from Sci Kit (sklearn) library are utilized.

Cleveland and Statlog datasets are utilized for preparing and testing purposes. In the experimentation, the coronary illness expectation is changed over into a binary model (i.e.) grouping into positive and negative classes. Along these lines, in the Cleveland dataset, the output 'Class' variable qualities are relegated 0 or 1. The past qualities like 2,3 and 4 are reassigned with the value of 1 which represents the nearness of coronary illness. For Statlog dataset since there are only two output variables 'Absent' and 'Present' the same is used as the class variables. Since the total out the number of records is just 303, 10-overlap cross approval is done in the tests to guarantee better outcomes. Gaussian Naive Bayes, Support Vector Machine, Decision Tree, Logistic Regression, and K Nearest Neighbours are utilized for the experimentation reason since these classifiers are demonstrated to best and customary ones.

4.1 Comparison Metrics

For contrasting the various classifiers' performance, the measurements e.g., such as exactness, accuracy, recall, and F1 scores are used. Equation 4, 5, 6, and 7 show the calculation of the previously mentioned performance measures. True positive and True Negative are the instances appropriately anticipated as positive and negative examples individually. False-positive instances are anticipated as positive though truly it is a negative example and the False Negatives are anticipated as negative samples however actually these are positive samples.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalseNegative + TrueNegative + FalsePositive} \tag{4}$$

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{5}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{6}$$

$$F1 = 2X \frac{Precision * Recall}{Precision + Recall} \tag{7}$$

Though all the four metrics seems to be same, they are different and used in different perspectives. Accuracy depicts how well our model in determining the true instances is. The precision determines the rate of predicting True positives in overall positives predicted (i.e.) True positives and False positives. Recall depicts the rate of True positives over real positive instances (i.e.) True positives and False Negatives. In certain application domain, a single metric is needed to make a judgement about the model and F1 is used there. F1 balances both Precision and Recall. F1 is particularly useful when uneven class distributions are there.

4.2 Individual Classifiers

The performance of individual classifiers in terms of accuracy, precision, recall and F1 scores are presented in Table 1 depicts for both Cleveland and Statlog dataset. Figure 4(a) corresponds to Cleveland and Figure 4(b) depicts the results for Statlog dataset. Naive Bayes shows better accuracy than all other classifiers followed by Support Vector Machine and Logistic Regression. Naïve Bayes gives better results than other individual classifiers because of its conditional independence of all the features. Since Naive Bayes, Support Vector Machine and Logistic regression are having accuracy more than 80% they are classified as Strong classifiers.

Considering Cleveland dataset, Naïve Bayes shows 24.32% and 11.07% improved accuracy than KNN and Decision Tree algorithms respectively. SVM classifier is 24.01% and 10.71% better than KNN and decision tree algorithms respectively. Logistic regression shows 22.80% and 9.28%improvement than KNN and decision tree algorithms respectively. The strong classifier algorithm like Naïve Bayes, SVM and Logistic regression is showing at least 20% more accuracy than KNN model and at least 10% more accuracy than decision tree algorithms. Since Decision Tree and K – Nearest Neighbour is having significantly lower accuracy they are classified as Weak classifier. KNN shows poor accuracy than all others because of its non-parametric approach.

Considering Statlog dataset, again Naïve Bayes is giving better results than all other individual algorithms. Naive Bayes is 20.26% and 9.25% better than KNN and decision trees respectively. Logistic regression follows next with 19.56% and 8.44% improved accuracy than KNN and decision tree algorithms. SVM shows an improvement of 19.20% and 8.04% than KNN and decision tree

Table 1. Comparison of Different Classifiers on Cleveland and Statlog Dataset

S. No	Classifiers	Cleveland				Statlog			
		Metrics				Metrics			
		Accuracy	Precision	Recall	F1	Accuracy	Preciion	Recall	F1
1	Logisti Regesson	0.8249	0.8421	0.7777	0.8022	0.8333	0.8234	0.8014	0.8070
2	NaivesByes	0.8415	0.8433	0.8054	0.8191	0.8407	0.8470	0.7833	0.8103
3	Decisio Tree	0.7483	0.7283	0.7325	0.7286	0.7629	0.7444	0.7270	0.7271
4	K Nearet Neighbors	0.6368	0.6172	0.5867	0.5937	0.6703	0.6470	0.5639	0.5989
5	SupportVector Machine	0.8381	0.8491	0.7878	0.8125	0.8296	0.8269	0.7858	0.8006

algorithms. In Statlog dataset also, it is observed than strong classifiers are around 20% more accurate than weak classifiers.

4.3 Classifiers With Bagging Technique

In the second experimental setup, the individual classifiers are enhanced using the bagging technique. Previously identified strong classifiers such as Gaussian Naïve Bayes, Support Vector Machine and

Figure 4. Accuracy of classifiers on Cleveland and Statlog Dataset

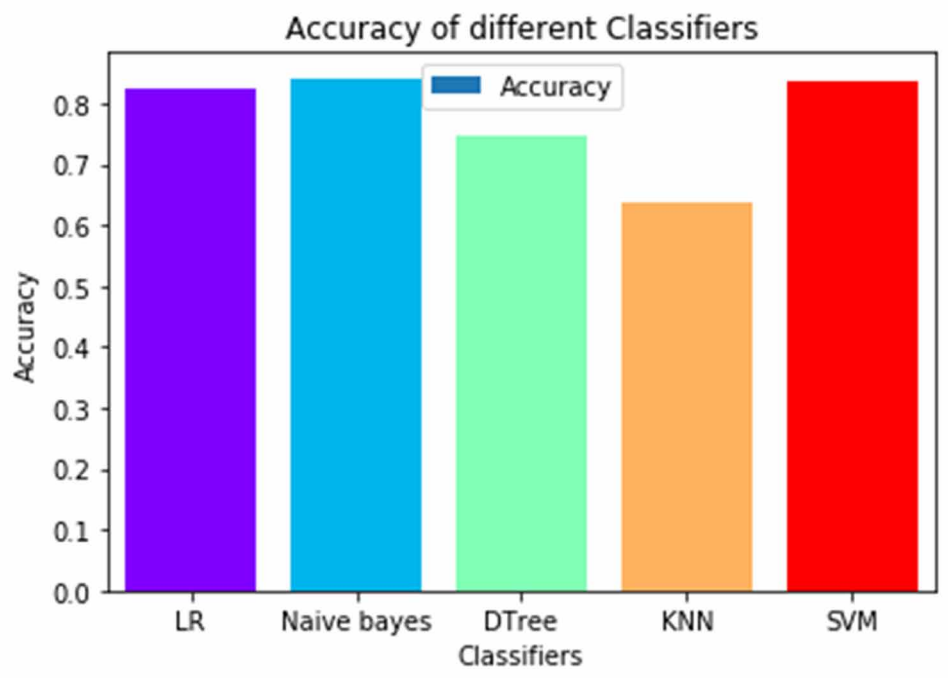


Table 2. Comparison of Different Classifiers using Bagging on Cleveland and Statlog Dataset

S. No	Classifiers	Cleveland				Statlog			
		Metrics				Metrics			
		Accuracy	Precision	Recall	F1	Accuracy	Preciion	Recall	F1
1	Logistic egresson	0.8249	0.8421	0.7777	0.8022	0.8333	0.8234	0.8014	0.8070
2	Gaussian Naives Bayes	0.8348	0.8355	0.8054	0.8179	0.8407	0.8408	0.7923	0.8120
3	Decision Tree	0.7982	0.7982	0.7675	0.7825	0.8222	0.8332	0.7717	0.7906
4	K Nearest Neighbors	0.6436	0.6285	0.5796	0.6029	0.6777	0.6487	0.5983	0.6177
5	Support Vector Machine	0.8282	0.8356	0.7927	0.8048	0.8370	0.8469	0.7858	0.8072

Logistic Regression again shows improved performance than KNN and Decision Trees. But there is no significant improvement in the performance of strong classifiers because of bagging technique. Whereas in the case of weak classifiers, Decision Tree accuracy increases from 74% to 79% for Cleveland dataset and 76% to 82% for Statlog dataset because of the bagging technique. In Cleveland, Naïve Bayes is 22.90% and 4.38% more accurate than KNN and decision tree algorithm. Bagging techniques work well for decision tree then all other classifiers. In Statlog, Naïve Bayes shows 19.38%

Figure 5. Classifiers Accuracy using Bagging for Cleveland Dataset

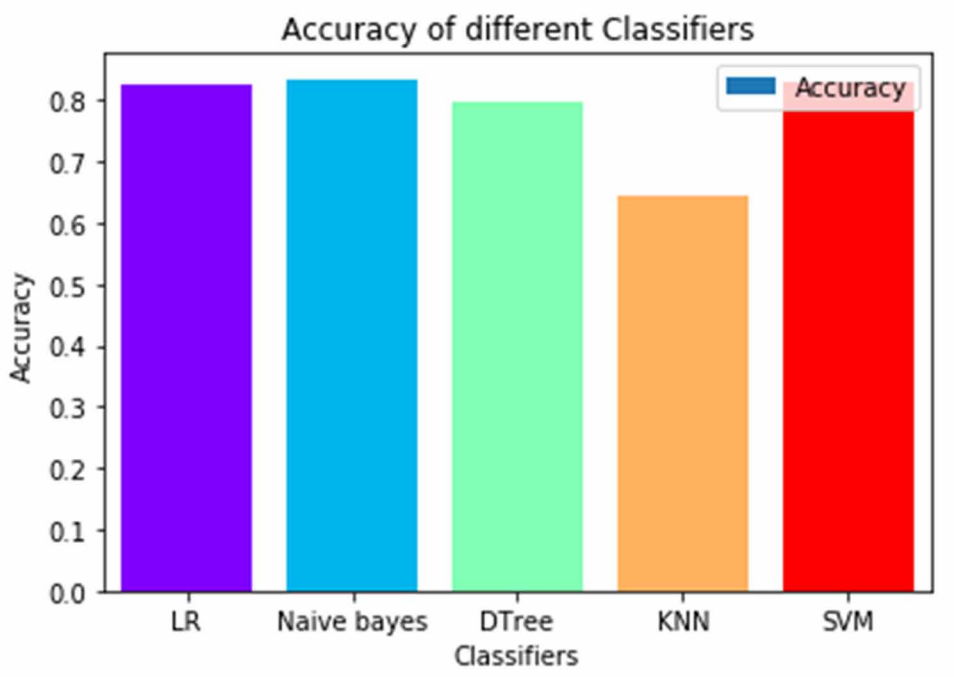


Table 3. Comparison of Different Classifiers using Boosting for Cleveland Dataset

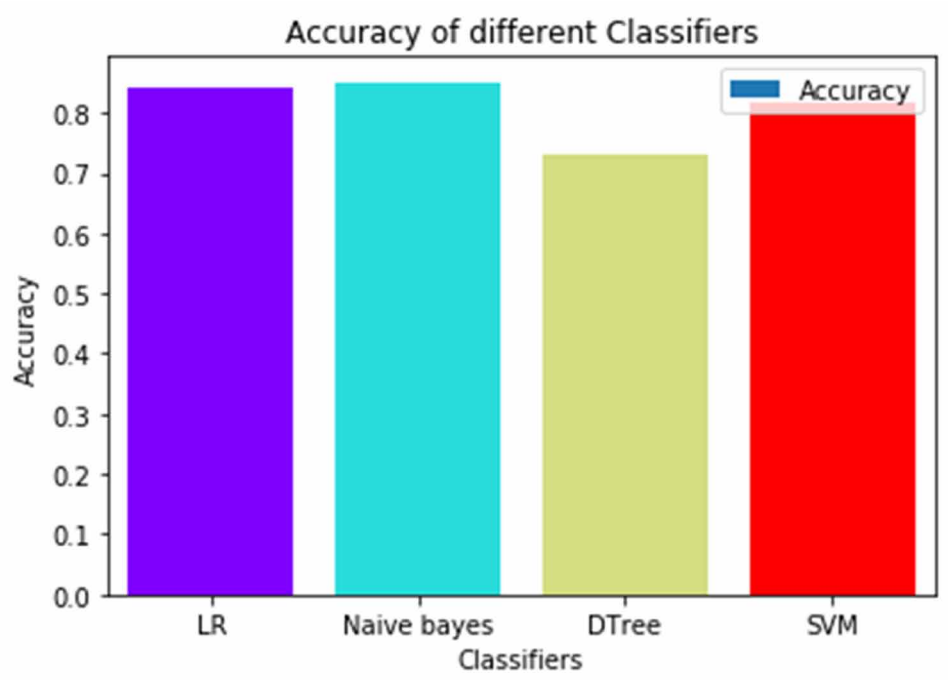
SS.No	Classifiers	Cleveland				Statlog			
		Metrics				Metrics			
		Accuracy (%)	Precision (%)	Recall (%)	F1	Accuracy	Preciion	Recall	F1
1	Logisti Regesson	0.8415	0.8433	0.8054	0.8238	0.8074	0.7942	0.7731	0.7779
2	Gaussia Naives Bayes	0.8513	0.8441	0.8321	0.8338	0.8481	0.8537	0.8004	0.8194
3	Decisio Tree	0.7317	0.7117	0.7084	0.7099	0.7481	0.7194	0.7400	0.7185
4	Support Vector Machine	0.8186	0.8080	0.7875	0.7924	0.8259	0.8527	0.7560	0.7875

and 2.20% improvement in accuracy than KNN and decision tree respectively. Table 2, Figure 5 (a) and 5 (b) depict the performance metrics of classifiers on Cleveland and Statlog datasets.

4.4 Classifiers With Boosting Technique

Table 3 depicts the result of Cleveland and Statlog dataset. Figure 6(a) depicts the comparison of individual classifiers enhanced with boosting technique for Cleveland dataset. Gaussian Naive Bayes shows highest accuracy of 85.13%, Logistic Regression at 84.15% and all others classifiers at the same

Figure 6. Classifiers Accuracy using Boosting on Cleveland and Statlog Dataset



performance level. Naïve Bayes improves accuracy by 14.04% than the decision tree algorithm. On comparing with other strong classifier algorithm SVM, Naïve Bayes shows 3.84% improved accuracy.

Figure 6(b) depicts the performance of classifiers using boosting techniques for Statlog dataset. In this also, Naïve Bayes is giving better results than all other classifiers. Naïve Bayes shows 11.79% improved accuracy than decision tree, 4.79% than logistic regression and 2.61% than SVM. Since KNN does not support sample weights, Boosting technique is not applied over KNN. In applying the boosting technique, Naïve Bayes is showing even considerable improvement than other strong classifier algorithms such as SVM and logistic regression.

Coming to our first research objective, Naïve Bayes as an individual classifier is giving the highest accuracy than all other classifiers. Also, Naïve Bayes enhanced using boosting technique shows the highest accuracy of 85.13% and 84.81% concerning Cleveland and Statlog datasets. It is because of the conditional independence of each attribute and its ability to get easily trained with even small datasets. In this research with the help of experimental results, it is established that naïve Bayes is the best classifier algorithm for predicting heart diseases.

Coming to our second research objective, (i.e.) the effect of bagging and boosting technique over weak classifiers, bagging technique improves the accuracy of decision tree algorithm by 6.66% for Cleveland and 7.77% for Statlog datasets. Bagging improves the accuracy of KNN by around just 1% for both Cleveland and Statlog datasets.

V. CONCLUSION AND FUTURE WORK

In this article, the base classifiers such as Logistic regression, Decision Tree, KNN, SVM and Naive Bayes are applied over Cleveland and Statlog data sets for heart disease prediction. The research objective was built on two aspects: (i) Identifying the best individual classifier and (ii) the effect of bagging and boosting on weak classifiers. Considering all the experimental results, Naive Bayes enhanced with boosting technique gives the highest accuracy of 85.13% and 84.81% for Cleveland and Statlog datasets. Also, it is identified that Naive Bayes, Support Vector Machine and Logistic Regression are strong classifiers with more than 80% accuracy and Decision Tree and K Nearest Neighbours as weak classifiers. Bagging and boosting techniques improve the performance of weak classifiers such as Decision Tree and K Nearest Neighbours. Bagging technique improved the accuracy of the decision tree algorithm by 7.77% maximum for Statlog dataset. In future, feature selection is to be applied to find out the most relevant features of the data set and applying over the ensemble models over it will give better-improved accuracy.

REFERENCES

Acharya, U., Faust, O., Sree, V., Swapna, G., Martis, R. J., Kadri, N. A., & Suri, J. S. (2014). Linear and nonlinear analysis of normal and CAD-affected heart rate signals. *Computer Methods and Programs in Biomedicine*, *113*(1), 55–68. doi:10.1016/j.cmpb.2013.08.017 PMID:24119391

Ahmad, G., Khan, M. A., Abbas, S., Athar, A., Khan, B. S., & Aslam, M. S. (2019). Automated Diagnosis of Hepatitis B Using Multilayer Mamdani Fuzzy Inference System. *Journal of Healthcare Engineering*, *2019*(2019), 6361318. doi:10.1155/2019/6361318 PMID:30867895

Alizadehsani, R., Abdar, M., Roshanzamir, M., Khosravi, A., Kebria, P. M., Khozeimeh, F., Nahavandi, S., Sarrafzadegan, N., & Acharya, U. R. (2019). Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Computers in Biology and Medicine*, *111*, 103346. doi:10.1016/j.compbiomed.2019.103346 PMID:31288140

Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., Bahadorian, B., & Sani, Z. A. (2013). A data mining approach for diagnosis of coronary artery disease. *Computer Methods and Programs in Biomedicine*, *111*(1), 52–61. doi:10.1016/j.cmpb.2013.03.004 PMID:23537611

Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics*, *36*, 82–93. doi:10.1016/j.tele.2018.11.007

Blake, C., & Merz, C. J. (2015). *UCI repository of machine learning databases*. Department of Information and Computer Science, University of California, Irvine, CA. http://www.archive.ics.uci.edu/ml

Cheung, N. (2001). *Machine learning techniques for medical analysis* (Doctoral dissertation). B. Sc. Thesis, School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, Australia.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, (pp. 1-15). Springer.

Indo-Asian News Service. (2018). *Microsoft and Apollo Hospitals Launch AI Model to Predict Heart Disease Risk.* Available Online: https://www.firstpost.com/tech/news-analysis/microsoft-and-apollo-hospitals-launch-ai-model-to-predict-heart-disease-risk-4988071.html

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (2015). *Machine Learning Repository, Heart Disease Data Set*. Available Online: https://archive.ics.uci.edu/ml/datasets/heart+Disease

Joloudari, Joloudari, Saadatfar, Ghasemigol, Razavi, Mosavi, Nabipour, Shamshirband, & Nadai. (2020). Coronary Artery Disease Diagnosis; Ranking the Significant Features Using a Random Trees Model. *International Journal of Environmental Health and Public Health*, *17*(731), 1-24. .10.3390/ijerph17030731

2Melillo, P., Izzo, R., Orrico, A., Scala, P., Attanasio, M., Mirra, M., De Luca, N., & Pecchia, L. (2015). Automatic prediction of cardiovascular and cerebrovascular events using heart rate variability analysis. *PLoS One*, *10*(3), e0118504. doi:10.1371/journal.pone.0118504 PMID:25793605

1Neshat, M., & Yaghobi, M. (2009). Designing a fuzzy expert system of diagnosing the hepatitis B intensity rate and comparing it with adaptive neural network fuzzy system. *Proceedings of World Congress on Engineering and Computer Science*, 797–802.

Palaniappan, S., & Awang, R. (2008). *Intelligent heart disease prediction system using data mining techniques. In IEEE/ACS international conference on computer systems and applications*. IEEE.

Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017). A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. In *IEEE Symposium on Computers and Communications (ISCC)*, (pp. 204-207). IEEE. doi:10.1109/ISCC.2017.8024530

Raghupathi, W. (2010). Data mining in health care. *Healthcare* Informatics. *Improving Efficiency and Productivity*, *211*, 223.

Roohallah, A., Abdar, M., Roshanzamir, M., Khosravi, A., Kebria, P. M., Khozeimeh, F., Nahavandi, S., Sarrafzadegan, N., & Acharya, U. R. (2019). Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Computers in Biology and Medicine*, *111*, 111–103346. doi:10.1016/j.compbiomed.2019.103346 PMID:31288140

Sardesai, A., Sambarey, P., Kharat, V., & Deshpande, A. (2014). Fuzzy logic application in gynecology: a case study. In *Proceedings of 2014 International Conference on Informatics,Electronics and Vision (ICIEV)*, (pp. 1–5). IEEE. doi:10.1109/ICIEV.2014.6850715

Srinivas, K., Raghavendra Rao, G., & Govardhan, A. (2010). Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. In *5th International Conference on Computer Science & Education*, (pp. 1344-1349). IEEE. doi:10.1109/ICCSE.2010.5593711

Uyar, K., & İlhan, A. (2017). Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. *Procedia Computer Science*, *120*, 588–593. doi:10.1016/j.procs.2017.11.283

Vincent, J. (2018). Google's new AI algorithm predicts heart disease by looking at your eyes. *The Verge*. Available Online: https://www.theverge.com/2018/2/19/17027902/google-verily-ai-algorithm-eye-scan-heart-disease-cardiovascular-risk

Wang, Y., Kung, L., Gupta, S., & Ozdemir, S. (2019). Leveraging big data analytics to improve quality of care in healthcare organizations: A configurational perspective. *British Journal of Management*, *2019*(30), 362–388. doi:10.1111/1467-8551.12332

World Health Organization. (2017). *Cardio Vascular Diseases*. Available Online https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Zipes, D. P., Libby, P., Bonow, R. O., Mann, D. L., & Tomaselli, G. F. (2018). *Braunwal's Heart Disease E-Book: A Textbook of Cardiovascular Medicine*. Elsevier Health Sciences Wiley.

Baranidharan Balakrishnan is currently working as Associate Professor, School of Computing, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India. He obtained his PhD in Wireless Sensor Networks from SASTRA University, M.Tech in Computer Science and Engineering from SRM University and B.E. in Electronics and Communication Engineering from Anna University. He has more than 20 International journal publication and 10 International conference publications. His area of interest includes WSN, IoT, AI, and ML.

C. N. S. Vinoth Kumar received Master of Engineering in Computer Science and completed his Doctoral degree from Annamalai University, India. His area of interest is Wireless Sensor Networks, Cryptography & Network Security, Artificial Intelligence, and Graphics. He is the lifelong member of Computer Society of India (CSI) and International associations of engineers. Currently He is working as Assistant professor, Department of Computer Science and Engineering in SRM Institute of Science and Technology, Chennai, Tamilnadu, India. His Current research includes Cyber Security related attacks and its approach using Machine learning algorithm.