



Prediction of Heart Disease with Machine Learning Algorithms

Rekha Pal

M.Phill, Research Scholar
Himalayan Garhwal University,
Pudhi Garhwal (UK), India
Email: drrekhapal@yahoo.co.in

Purnima Tyagi

Associate Professor,
Department of Computer Applications
Himalayan Garhwal University,
Pudhi Garhwal (UK), India
Email: poornima.tyagi@gmail.com

ABSTRACT

Coronary illness (Heart Disease) is one of the most widely recognized sicknesses all through the world. In this article, we depict an inventive AI framework and group procedure that can precisely distinguish HD and apply it to the data gathered from the UCI AI store. We attempted six conventional AI calculations (SVM, DT, KNN, RF, Adaboost and LR), and two basic classifiers (SVC, KNN, DT, RF, ADA and SVC, KNN, RF) as ensemble calculation. To work on the portrayal and correlation of these calculations, the information was normalized. The introduced procedure works on the depiction of all customary AI estimations used in this outline. We propose voting classifiers 1 and 2 as group calculations for AI calculations. Our exploration shows that RF makes an exactness of 88.52% in a bunch of AI classifiers, while Voting Classifier gives a precision of 86.88%, while anticipating that it will be recorded as a notable UCI AI dataset. We reasoned that AI programs overhauled through the proposed techniques can recommend surprisingly exact models that are gotten ready for clinical use and assessment.

Keywords:— Heart disease (HD), Ensemble technique, UCI machine learning repository, voting classifiers, Mortality, Random forest.

I. INTRODUCTION

Coronary illness is the most widely recognized cardiovascular disease. It is the principal reason of worldwide demise [1]. World Health Organization (WHO) found that around 17.8 million passed on from CVD in 2017, while 330 million people lost their life till now, and 35.6 million people experiencing coronary sickness [2, 3]. To put it plainly, exact CAD affirmation and proper treatment can keep away from numerous CAD-related illnesses and passings. This examination means to introduce another cycle (a mix of a couple of AI estimations and outfit computations) to distinguish CAD. Data mining is a PC based cycle used to remove critical information from a great deal of outstanding data. Some amazing data mining computations include: SVM, NN, DT, Genetic calculations and Bayesian Networks (BN). All of these progressions enjoys the two benefits and deficiencies and should be used with alert. Contingent upon the kind of data and the condition of examination (for instance, every calculation incorporates variables and limits), a portion of these systems can acquire gigantic outcomes anticipating sicknesses, while different procedures may not accomplish the ideal outcomes. The portrayal of innovation relies upon the possibility of the calculation and the sort and execution status

of the information, which may truly influence the outcomes. Thusly, changing the procedure precisely to the kind of data to be mined and observing the ideal outcomes can work on the introduction of the innovation.

Nowadays, the utilization of different AI calculations in various applications is step by step well known and typical. Since how much data connected with these various applications is gigantic and multifaceted, it is critical to eliminate covered up and important information from them [4]. This information can be utilized to work on the idea of various specialists [5]. Disaggregating these information, affiliations, legislatures, experts and other staff can give further developed kinds of help and increment the worth of coordinated effort with clients, patients and others. This examination centers around the utilization of AI and information mining advancements to give clinical assistance information [6]. In the field of clinical and clinical organizations, surprising AI frameworks have hardly been used for a broad investigation of season of sicknesses, similar to, Parkinson's, liver, coronary, bosom, lung related infection, etc. As a general rule, the applied innovation has gotten incredible outcomes expecting explicit infections. The simple assurance of these illnesses (counting growths) is the most unfathomable. Obviously, the disclosure cycle requires uncommon data and experience. We perceive AI, since we can utilize data mining to work on the exactness of procedures, decrease the quantity of demonstrative blunders, and eventually pass quality administration to patients [7].

The principle objective of this article is to plan to utilize a few AI and data mining strategies to recognize new and convincing models of CAD [8]. In this exploration, we used the scientific emotionally supportive

network as a clinical choice emotionally supportive network subject to its commonness and wide application in various data structures. SVM, DT, KNN, RF, Adaboost and LR and two democratic classifiers (SVC, KNN, DT, RF, ADA) and (SVC, KNN, RF) are utilized for this test. At long last, the two estimations that played out the best were chosen (classifiers SVC, KNN, and RF, abridged as casting a ballot classifier 2). This clinical Cardio Artery Disease (CAD) informational collection utilized in our assessment included 303 records and 14 highlights.

The leftover work is as per the following. Segment 2 presents the methodology. Segment 3 covers the dataset information. Section 4 contains the results and conversation from the investigation. The conclusion from results is discussed in Section 5.

II. METHODOLOGY

Six machine learning algorithms were included in this research, one of which performed better overall and also applied ensemble techniques (voting classifier 1 and voting classifier 2). The voting classifier 2 performed better in this group [9]. The following is the combination of classifiers in the ensemble technique.

1. Voting classifier 1 (SVC, KNN, DT, RF, ADA),
2. Voting classifier 2 (SVC, KNN, RF).

These strategies are utilized to improve the presentation of AI algorithms. Accordingly, an information normalization technique was applied, which is utilized as a pre-preparing strategy for the development strategy. Figure 1 shows the proposed framework of the model. The following introduces the selected machine learning classifier and ensemble classifier.

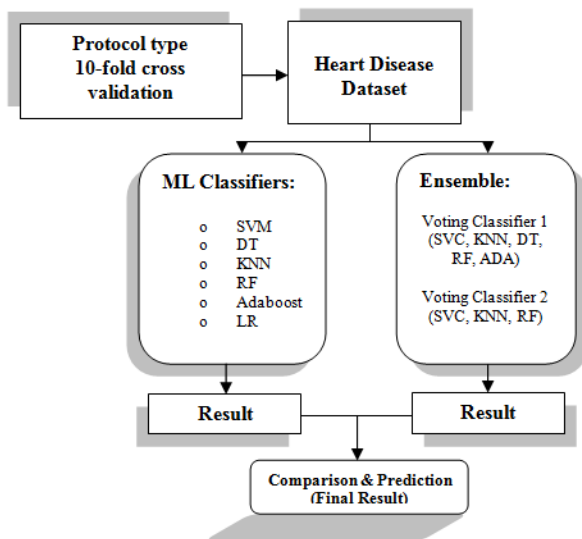


Figure 1 : Proposed model for prediction of Heart Disease

Ensemble Technique

The ensemble strategy uses a large number of learning calculations to obtain desirable, foresighted execution, and this can only be obtained from any constituent learning calculations. Compared with the measurable equipment in the usually endless factual mechanics, AI ensemble only includes a limited and real arrangement of elective models, but it is generally believed that among these other options, there are substantially more adaptive structures [10].

In this paper, two voting classifiers are used as the ensemble technique. In the first voting classifier, SVC, KNN, DT, RF and AdaBoost are combined; in other voting classifiers, SVC, KNN and RF are combined.

Feature Correlation

The correlation feature is the ratio of the direct connection between two quantitative factors, similar to height and weight. We can also describe the relationship as the ratio of the dependence of one variable on another [11].

High correlation is usually a valuable attribute-if two factors correspond deeply, we can difference the other. Therefore, we generally look for correlated features that are particularly relevant to the goal, especially for direct AI models.

Nevertheless, if there is an abnormal connection between the two factors, they will provide too much data about the target. Fundamentally, we can select foresee the objective utilizing just one of the repetitive components.

Confusion Matrix

To check the portrayal of the classifier, diverse execution assessment measurements are utilized in this investigation [12]. From the confusion matrix, we find the following:

Table 1: Confusion Matrix

	Heart Disease (YES)	Heart Disease (NO)
Have heart disease (YES)	TP	FN
Have heart disease (NO)	FP	TN

Accuracy: It quantifies the number of positive and negative opinions that are effectively grouped.

$$\text{Accuracy} = \frac{TP + TN}{(TP) + (TN) + (FP) + (FN)} * 100$$

III. DATASET ANALYSIS

Many researchers and experts have analyzed the heart disease data set, which can be obtained from the UCI machine learning online repository [13]. In this study, this data classification was used to plan cardiovascular disease structures that depend on AI. The model size of the UCI

Heart Disease Data Set Information Index is 303 patients with 14 features and no missing values. There are two target categories of patients with coronary artery disease or patients without heart disease. Figure 2 lists all the data and depicts 303 instances of the 14 features of the data list.

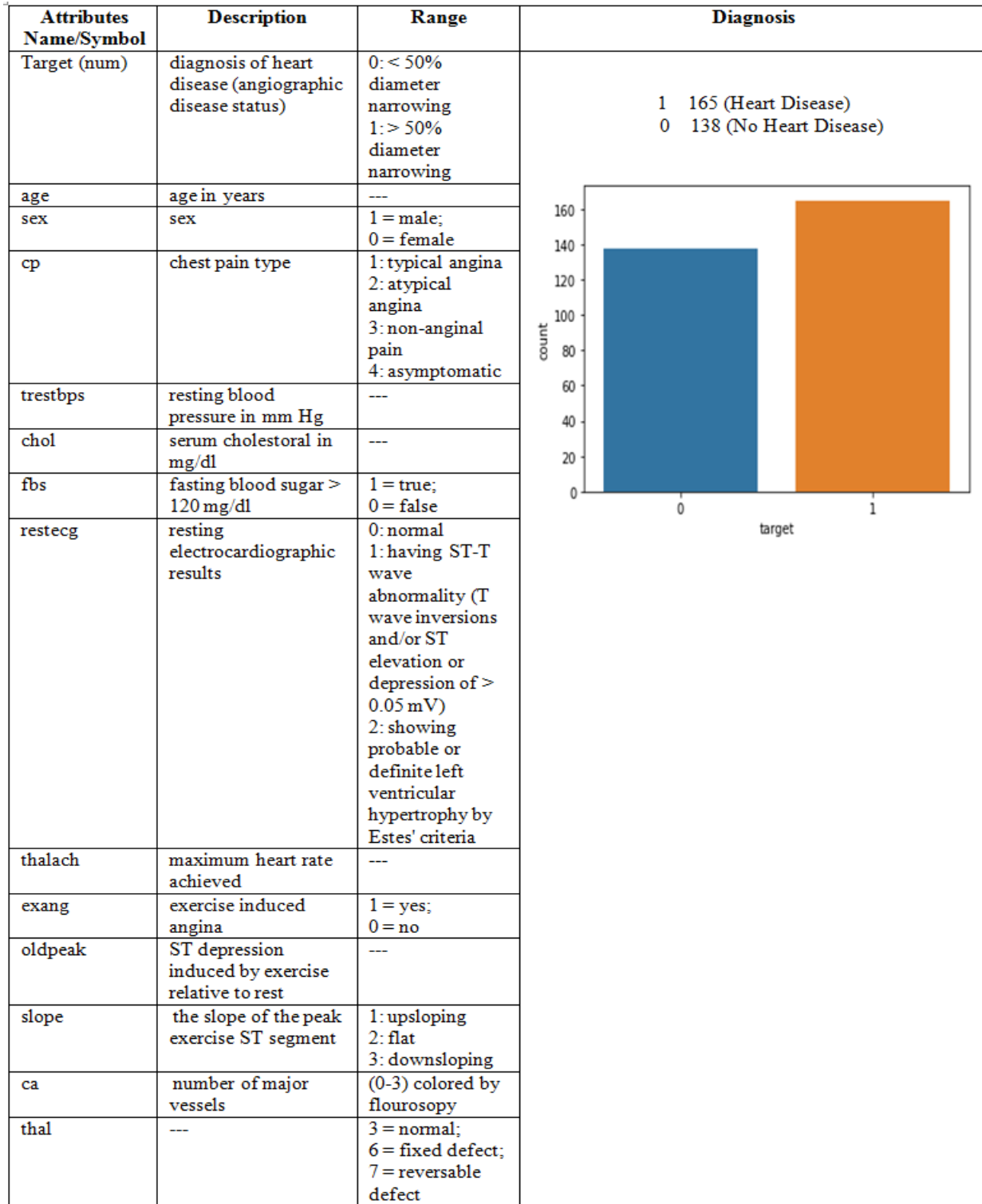


Figure 2: Attributes of Heart Disease

IV. RESULTS AND DISCUSSION

Each attributes and their numeric values provide help in disease prediction. By the help of box and whisker, we have implemented the heart disease attributes in brief and measured each attributes distribution. Figure 3 shows the box and whisker plot of the data set attributes.

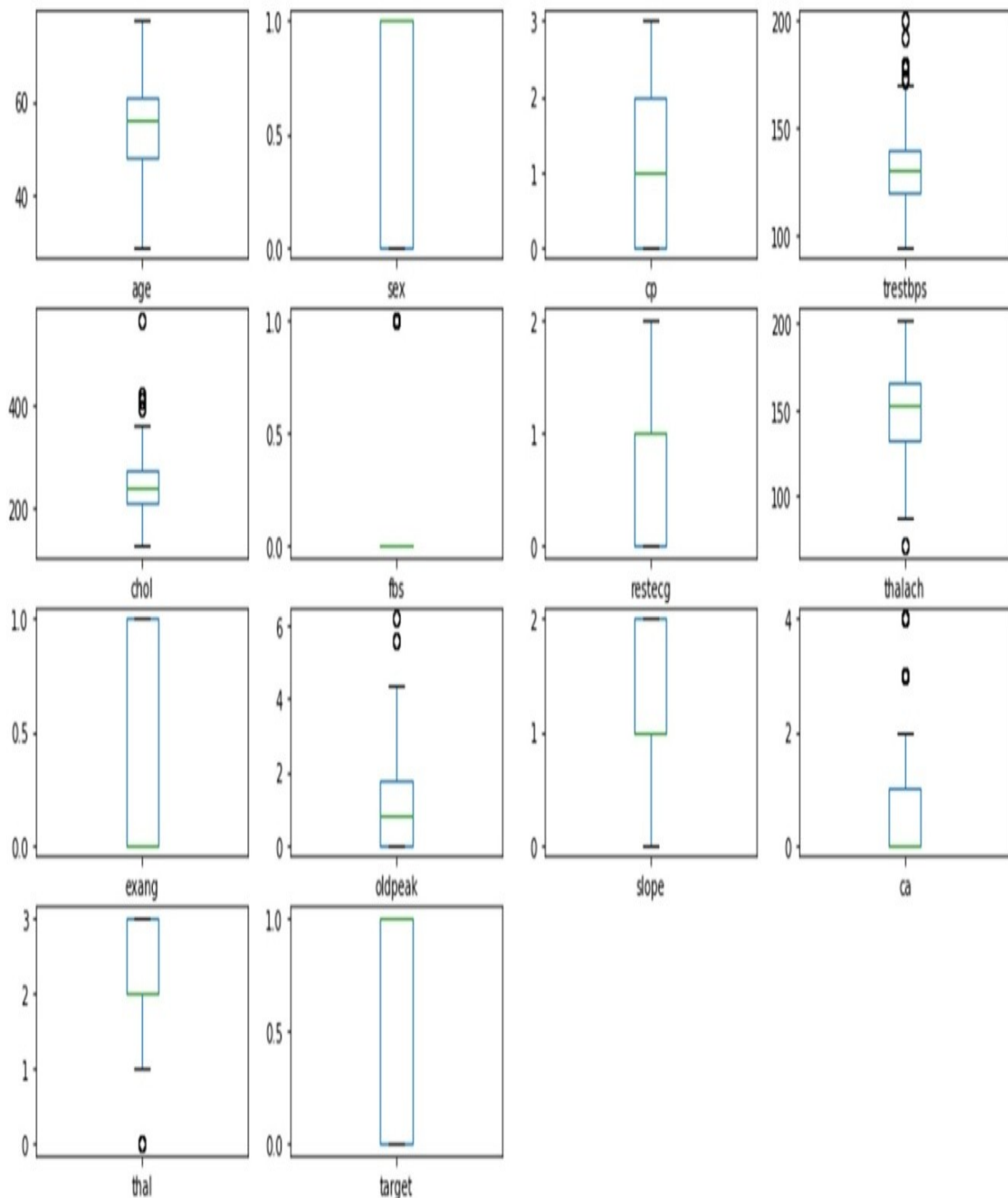


Figure 3: Representation of Box and Whisker plotting of heart disease attributes

Similar checks for different classification algorithms have been performed on the heart disease data set. Some algorithms showed high accuracy, while results of some algorithms were inadequately. In order to improve the display of weak classifiers, ensemble technique is used. This work uses ensemble techniques, such as voting. The voting classifier uses SVM, DT, KNN and RF as voting 1 and SVM, KNN and RF as voting 2.

In this section, we discuss the findings of the experiment. As mentioned earlier, in the main test, six calculations were checked, including SVM, DT, KNN, RF, Adaboost and LR. The best in the rest of our exams, the classifier (RF) was chosen. A summary of our survey usage is as follows:

The correlation of features is shown in Figure 4 there are two types of features that are positively and negatively related to the target feature. Compared with restecg, Cp, thalach, and slope are highly positively correlated features [14]. It can also be seen that exang, oldpeak, gender, thal, fbs, chol, ca, age and trtbps are negatively correlated with the target variable. Through this Figure 4, it can be concluded that attributes have their importance to improve the accuracy of the classifier.

Table 2 and Figure 5 show the accuracy results of the ML classifier and the overall results. Among the six machine learning classifiers, the accuracy of the random forest test data set is 88.52%, which is better. The accuracy of other classifiers is

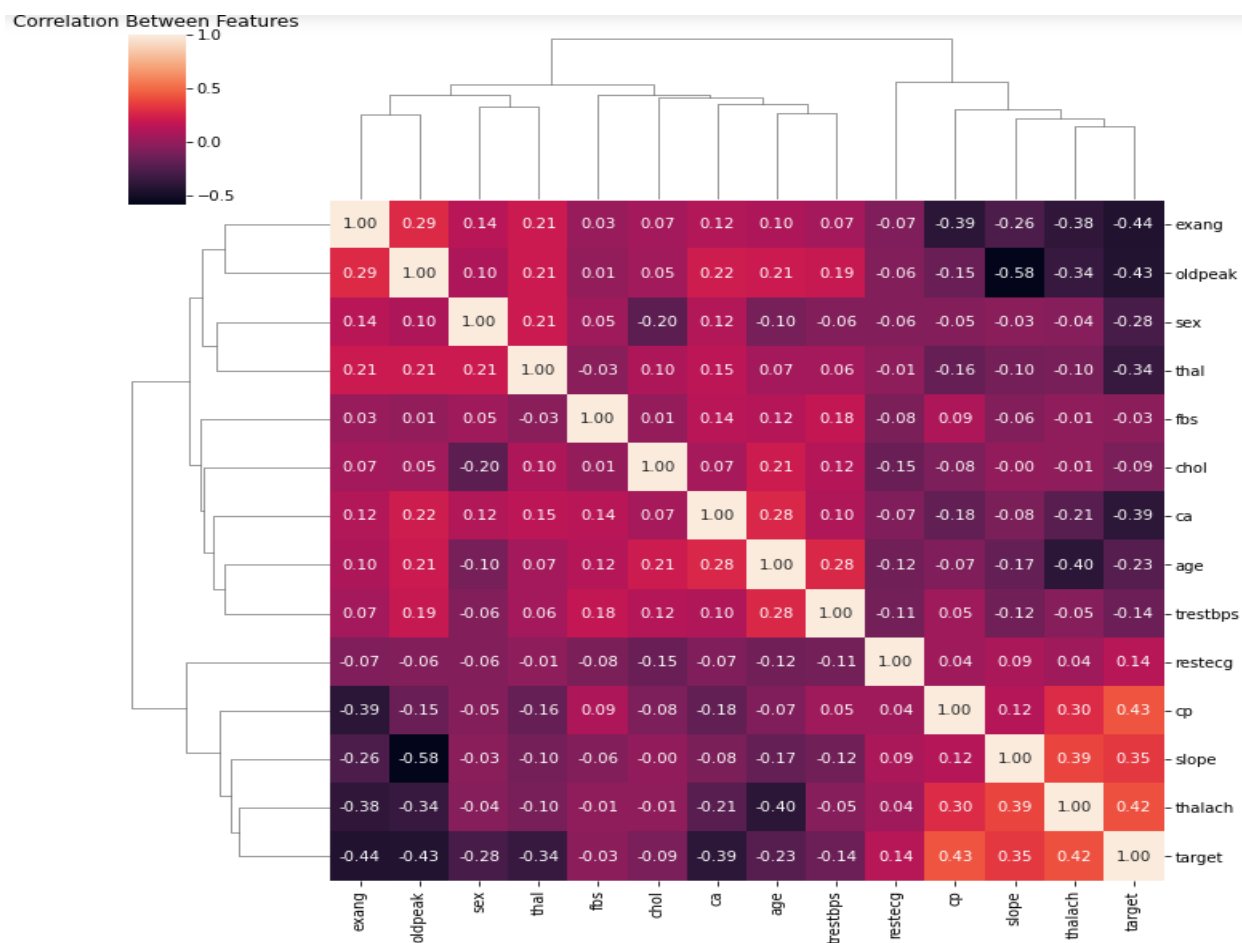
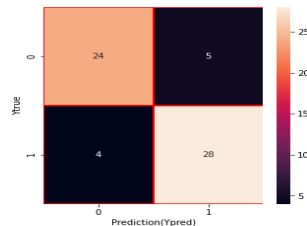
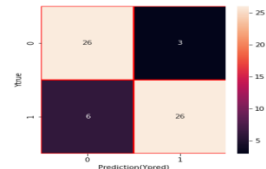
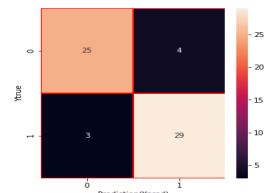
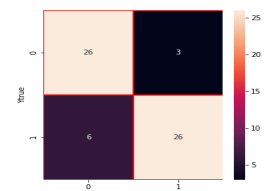
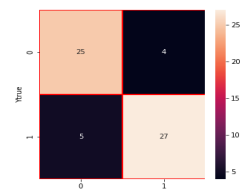


Figure 4: Features Correlation

the same, that is, 85.24%, while the accuracy of KNN is 86.88%. Therefore, random forest is a classifier suitable for heart disease data set prediction [15].

Table 2: Results obtain by ensemble and ML classifiers

Result of Machine Learning Classifiers		
Classifier Name	% Accuracy	Confusion Matrix
KNN	Test: 86.88 Train: 88.84	[[25 4] [4 28]]
SVC	Test: 85.24 Train: 91.32	
DT	Test: 85.24 Train: 88.42	
RF	Test: 88.52 Train: 87.60	
AdaBoost	Test: 85.24 Train: 88.42	
LR	Test: 85.24 Train: 86.36	
Result of Ensemble		
Voting Classifiers 1 (svc, knn, dt, rf, ada)	Test: 85.24 Train: 91.32	
Voting Classifiers 2 (svc, knn, rf)	Test: 86.88 Train: 90.49	

In Table 2 above, it is concluded that the accuracy of the results of the ensemble voting classifier 1 (SVC, KNN, DT, RF and Adaboost) is 85.24%, which is less than the accuracy of the voting classifier 2 (SVC, KNN and RF), which is 86.88%. Therefore, in prediction of heart disease, the ensemble result of voting 2 is better than voting 1.

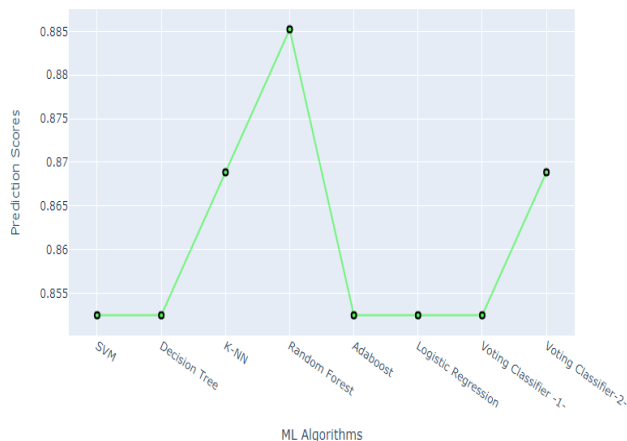


Figure 5: Performance of the classifiers

V. CONCLUSIONS

This article uses a combination of classifiers to investigate the accuracy of expected coronary artery disease. The heart informational collection from UCI AI vault is utilized for preparing and testing purposes. This examination utilizes ensemble strategy with support vector machine (SVM), decision tree (DT), k nearest neighbor (KNN), random forest (RF) and Adaboost. Combine these total calculations into 2 vote classifiers to generate vote 1 (SVM, KNN, DT, RF, and AdaBoost), while other sets include (SVM, KNN, and RF) as vote 2. Nevertheless, the six machine learning classifiers are also applied to the data set. In the entire classifier, random forest has the highest accuracy, such as 88.52%. Examination of the results shows that RF is most worthy of attention in terms of accuracy.

REFERENCES:

- [1] Wah, T. Y., Gopal Raj, R., & Iqbal, U. (2018). Automated diagnosis of coronary artery disease: a review and workflow. *Cardiology research and practice*, 2018.
- [2] Zainel, A. J. A. L., Al Nuaimi, A. S., & Syed, M. A. (2020). risk factors associated with cardiovascular diseases among adults attending the primary health care centers in qatar a cross sectional study. *Journal of Community Medicine & Public Health*.
- [3] Hu, K. Prevalence and Challenges of Hypertensive Heart Diseases in the Real World.
- [4] Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. *Jama*, 309(13), 1351-1352.
- [5] Parikh, R. B., Kakad, M., & Bates, D. W. (2016). Integrating predictive analytics into high-value care: the dawn of precision delivery. *Jama*, 315 (7), 651-652.
- [6] Darcy, A. M., Louie, A. K., & Roberts, L. W. (2016). Machine learning and the profession of medicine. *Jama*, 315(6), 551-552.
- [7] Nath, S. V. (2006, December). Crime pattern detection using data mining. In 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops (pp. 41-44). IEEE.
- [8] Cottrell, J. A., Hughes, T. J., & Bazilevs, Y. (2009). Isogeometric analysis: toward integration of CAD and FEA. John Wiley & Sons.

- [9] Anbarasi, M., Anupriya, E., & Iyengar, N. C. S. N. (2010). Enhanced prediction of heart disease with feature subset selection using genetic algorithm. *International Journal of Engineering Science and Technology*, 2(10), 5370-5376.
- [10] Khemphila, A., & Boonjing, V. (2011, August). Heart disease classification using neural network and feature selection. In *2011 21st International Conference on Systems Engineering* (pp. 406-409). IEEE.
- [11] Mokeddem, S., Atmani, B., & Mokaddem, M. (2013). Supervised feature selection for diagnosis of coronary artery disease based on genetic algorithm. *arXiv preprint arXiv:1305.6046*.
- [12] Wisaeng, K. (2014). Predict the diagnosis of heart disease using feature selection and k-nearest neighbor algorithm. *Applied Mathematical Sciences*, 8(83), 4103-4113.
- [13] Jabbar, M. A. (2017). Prediction of heart disease using k-nearest neighbor and particle swarm optimization.
- [14] Vijayashree, J., & Sultana, H. P. (2018). A machine learning framework for feature selection in heart disease classification using improved particle swarm optimization with support vector machine classifier. *Programming and Computer Software*, 44(6), 388-397.
- [15] Reddy, G. T., Reddy, M. P. K., Lakshmana, K., Rajput, D. S., Kaluri, R., & Srivastava, G. (2020). Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis. *Evolutionary Intelligence*, 13(2), 185-196.

* * * * *