Volume 11, Issue 1
February 2022

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

AN ANALYSIS OF MACHINE LEARNING TECHNIQUES IN EARLY DETECTION OF HEART DISEASE

[1]Bhabesh Deka and [2]Bismita Choudhury

[1]Research Scholar, Dept. of Computer Science and Engineering, Assam down town University, India
[2]Assistant Professor, Dept. of Computer Science and Engineering, Assam down town University, India
E-Mail: [1]vincy.ghy@gmail.com, [2]bismi.choudhury@gmail.com

ABSTRACT

Heart disease has become one of the primary concerns for the healthcare sector. The rapid growth of heart-related problems in people needs to be controlled throughout the globe. At present almost all medical sectors maintain their patient information for their interest and the use of data mining techniques in medical databases can be very useful to predict disease at an early stage. There are many data mining algorithms available that can be applied to analyze unknown and hidden facts in a database. To achieve optimum output the selection of proper data mining techniques is very essential and it is also difficult to guess which algorithm will perform better in a particular situation. This research aims to review a comparison of different important data mining algorithms used by various researchers in their experiments especially in the prediction of heart disease. Also, a model is proposed that includes traditional techniques as well as ensemble techniques to compare their performance in the early detection of heart disease. The ensemble of five different classifiers Naïve Bayes, Logistic regression, Decision tree, k-Nearest Neighbor (kNN), and Support Vector Machine (SVM) showed a maximum accuracy of 92.1% in our experiment.

Keywords: *Data mining, Disease Prediction, Machine Learning, Python, Ensemble Technique*

1. INTRODUCTION

Cardiovascular disease (CVD) is one of the most popular causes of death across the world. People of any age group can be affected by this disease. In general, heart disease indicates several conditions of the heart where its working ability decreases due to different reasons. There are so many risk factors that trigger heart disease, like unhealthy food habits, smoking, uncontrolled blood pressure, high level of bad cholesterol, physical inactiveness, obesity, etc. As mentioned by World Health Organization (WHO), due to CVD approximately 179 lakh people lose their lives every year, and that is around 31% of all deaths globally. It has become a prime cause of morbidity and mortality in today's modern society [7]. As published by a web portal (downtoearth.org.in) the prevention of cardiovascular disease is one of the significant needs among India's sustainable improvement objectives in light of the fact that the disease influences the nation's working population. Also, the MCCD (Medical Certification of Cause of Death) report of 2018, around 57% of the total deaths were because of CVD among people of age 25 to 69 years. To overcome the problem of increasing heart disease patients in the entire world there must be some technique to detect the risk of CVD at an early stage. In this regard, the use of data mining can be very beneficial for the health care sector. Data mining is primarily a process to analyze data and is designed to explore meaning full information within a large dataset, such as market or business-related. It searches different variables and finds out the existence of consistent patterns as well as the systematic relationship between those variables. The main aim of data mining is prediction and that is why most business applications use predictive data mining.

The use of data mining in the prediction of any disease can play an important role as it can explore the various hidden relationships between data in a data warehouse. The data mining approach can be applied to predict so many diseases such as cancer in the lungs, thyroid disease, breast cancer, diabetes, hepatitis, liver cancer, etc. Due to the improvement of technology, access to medical data is growing day by day, hence healthcare organizations are now concentrating on how to improve the quality of their patient data to optimize the efficiency so that data mining can be used. Data mining has been found effective in many areas of the healthcare industry such as relationship management of customers, predictive medicine, identification of fraud and abuse, management of healthcare, and also measuring the effectiveness of different treatments [17]. While detecting a disease, the patient needs to undergo several tests which is time-consuming and also cost-effective. This can be minimized by using effective data mining techniques with optimum accuracy. According to many data experts, the overall expense cost of health care can be reduced by almost 30% by the proper application of data mining [18]. In this study, an analysis of different data mining algorithms is presented based on performance in terms of accuracy and efficiency, especially in heart disease prediction. This research aims to check which type of algorithm performs better results by studying different research experiments through machine learning techniques (MLT) in the healthcare sector. Here initially, we try to check the performance of individual base classifiers that are used by various researchers by comparing their results. Also, this study focuses on the use of ensemble techniques to improve the performance of weak classifiers for which a model is proposed in fig-3.

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

It is also mentioned by many researchers in their experiment that ensemble technique is a better idea for improvement of model performance especially in the prediction of disease. The proposed approach includes three separate experiments to check the accuracy levels of different classifiers individually and also by making their ensemble. In our experiment, we have implemented various ML techniques in the heart disease dataset with the help of the Python machine learning tool.

2. REVIEW OF LITERATURE

Extracting useful information from any huge database can be very complex and it needs smart algorithms. For this, different research activities have been carried out around the world to find out the best data mining algorithm with optimum accuracy and efficiency. But it is very difficult to guess which algorithm will perform best in a particular situation. Through this paper, some research activities related to cardiovascular disease (CVD) prediction through machine learning algorithms have been studied to compare their performance in different situations. A few papers not related to CVD are also considered to check the performance of machine learning classifiers in the case of other diseases.

V.Kirubha and S.Manju Priya in their research paper presented an analysis that compares the performance of different algorithms in disease prediction especially Heart, Liver, Kidney, Diabetes, and Cancer disease. They observed that in the diagnosis of different diseases the final result may vary concerning the tools and technology used. They also mentioned that data mining can show better results in disease prediction if appropriate tools and techniques are applied [2]. Priyanka P. Shinde et. al presented a paper to analyze data mining techniques to predict heart disease in aggregation with a real-time dataset. They found that in many studies, a Genetic algorithm performs an important role in the prediction of heart disease and can be combined with other data mining techniques to provide the best accuracy. They also found that people in the age group of 40 to 60 should be aware as this group has a higher probability of heart disease [3]. Meenu Singla and Kawaljeet Singh in their research paper presented a system to predict heart disease using data mining clustering technique. Their main objective was to predict the presence of heart disease with more accuracy but in very little time. In their experimental work, they have used the Weka tool with Pima diabetes data set to check the performance of their proposed algorithm. In the final result, they found that the Farthest first clustering algorithm performed best in comparison with other algorithms and their system has the provision for further expansion [4]. Amit Tate et al presented a paper to compare data mining classification algorithms in the prediction of disease. They have used the WEKA tool with a high dimensional dataset in their experiment and found that the Random forest (Ensemble classifier) algorithm had performed well considering all possible factors [5]. L. Parthiban and R. Subramanian presented an intelligent system that can predict heart disease using CANFIS and genetic algorithms. They have tested their prototype with a heart disease training dataset collected from the University of California and found that the Genetic algorithm is very effective in the automatic tuning of CANFIS parameters and best in the selection of feature sets [6]. L. Abdullah, in his paper, presented a model using the Fuzzy linear regression technique for identifying risk factors of coronary heart disease. The main objective of this study was to assess various risk factors that can contribute to heart disease. He tested 130 patient data collected from govt. hospital from Malaysia using the matrix-driven multivariate fuzzy linear regression model. The experimental result was good inaccuracy [7]. Durga Kinge and S. K. Gaikwad in their research showed the efficiency of classification algorithms using the WEKA tool. They tested their experiment with seven algorithms including ensemble techniques (Bagging and Boosting) with UCI Heart-disease dataset that consists of 303 instances. Finally, in their experiment, logistic regression, random forest, and Naive Bayes algorithms performed very well in heart disease prediction [8]. K. Aparna et al proposed hybrid data mining techniques to predict heart disease. In the proposed prototype (Intelligent Heart Disease Prediction system) they used three data mining models, which are Decision tree, Naïve Bayes, and Neural Network. Their data source consists of 909 records with 15 attributes and the result shown was good for all three models [9]. M.A. Nishara Banu and B. Gomathy through their research work used data mining classification algorithms to predict heart disease. They mainly used the MAFIA (Maximum Frequent Itemset Algorithm) for data classification. Their result shows an accuracy level up to 94% [10].

C. Latha and S. C. Jeeva in their paper completed a comparative analytical analysis to determine whether ensemble techniques can be employed for the improvement of heart disease prediction accuracy. Their aim was not only to improve the accuracy of the weak classifier but also its implementation with a medical dataset to check its prediction ability of a disease at an early stage. In their experiment, they used the Cleveland heart dataset, which consists of 14 attributes of 303 patient records. After successful implementation of their model, the result shows that ensemble techniques, bagging or/and boosting, are efficient in enhancing the accuracy of heart disease prediction and identifying the risk of factors. Using ensemble techniques, the accuracy of their model increased by 7%. In their study, the 'majority voting' algorithm obtained the highest improvement in CVD prediction accuracy [20]. I. D. Mienye et al in their study, proposed an improved machine learning method for predicting the risk of heart disease. Their proposed technique partitioned the training dataset into smaller subsets with the help of mean based splitting method. Each subset of the datasets was modelled using a “Classification And Regression Tree” (CART). They applied an uniform ensemble technique

with the help of accuracy-based weighted aging classifier. Their technique is a modified version of WAE (weighted aging classifier). After successful implementation of their experiment in Python machine learning environment for both Cleveland and Framingham heart disease datasets, they obtained maximum accuracy of 93% and 91% respectively [21]. Anurag Kumar Vermaa et al in their research two experiments conducted for the prediction of skin disease and they have taken the UCI Skin disease (University of California Irvine) dataset, which consist of 34 attributes of 366 samples. In their first experiment, they used three ensemble methods such as Bagging, Gradient boosting, and AdaBoost on six different classifiers considering all 34 features available in the dataset. The selected six classifiers are Linear Discriminant Analysis, Passive Aggressive Classifier, Gaussian Naïve Bayesian, Bernoulli Naïve Bayesian, Radius Neighbors Classifier (RNC), and Extra Tree Classifier. Again in the second experiment, they used the feature selection (FS) method and selected only 15 important attributes from the dataset. In both experiments, the Gradient boosting ensemble technique generated the highest accuracy of 99.46% (without FS) and 99.68% (with FS) when applied on RNC [22].

In this study, we have considered the approaches related to heart disease prediction but in some activities researchers also attempted to check their experiment in different diseases like kidney, diabetes, Liver disease, etc. In many experiments, genetic algorithms are used to predict the risk of heart disease with the UCI machine learning dataset. To improve the accuracy of prediction ability a few researchers proposed a hybrid technique that combines different weak classifiers to produce more accurate results. This approach is also known as the ensemble of classifiers. To build a strong model for disease prediction it is suggested by many researchers that the ensemble approach can play a better role as compared with the traditional base classifiers. Also, in many research activities, we have seen that a proper feature selection algorithm has a great effect in improving model performance because too many irrelevant parameters can increase the noise in the ML process.

3. COMPARATIVE ANALYSIS

In our study, we have collected the accuracy levels of different data mining algorithms from different research papers. Many algorithms are applied to achieve the best result with the different environments especially with medical datasets. We have seen that various algorithms have shown a different level of accuracy with different data items. The accuracy levels against the algorithms with paper numbers are shown clearly in the following table (Table-1).

From the above table, it is seen that most of the researchers used Weka and Python machine learning environment for the analysis of the medical dataset. Depending on the algorithms used it is found that the accuracy level ranges within 50% to 99% approximately. It is also observed from the above comparison that in some cases the original model also performed better results than the ensemble technique. But in major cases where both original and ensemble classifiers are used, the researcher found that the ensemble approach can perform much better in terms of efficiency and accuracy. To compare the same, this study proposes a technique for early detection of cardiovascular disease risk by using both ensemble and original classifiers. After analyzing different previous research activities, we have chosen a few well-known classifiers for our experiment in three different ways. These include Naïve Bayes (NB), k-NN, Logistic Regression (LR), Decision tree (DT), Random Forest (RF), and Support Vector Machine (SVM). A heterogeneous ensemble will be created by using five classifiers and also Random Forest will be used as a homogenous ensemble technique for this study. The primary aim of this research is to check which technique can predict heart disease in a patient more accurately. The following figure (Fig-1) shows the proposed model architecture with different levels.

4. THE ENSEMBLE APPROACH

The basic idea behind the ensemble method is that it uses multiple similar or different independent models (weak learners) to generate an output of some kind of prediction. Through this approach, a machine can generate a better predictive performance as compared to a single base classifier. The ensemble approach is a very effective Meta classification technology that aggregates the weak classifiers with strong learners to improve the overall efficacy of the weak classifiers [20].

It has been seen that in most of the cases ensemble technique performs better in comparison to the original classifier as it combines various models into one usually more accurate than the best of its components. In many situations, the ensemble approach tries to improve the result of the machine learning system by combining different weak learners like the random forest, SVM, Logistic regression, etc. and this approach is also called a heterogeneous ensemble. On the other hand, when an ensemble is done with a set of classifiers of the same type is called a homogenous ensemble. For example, the Random forest model is an ensemble of multiple decision trees. Here, it is most important to note that the datasets should be sampled separately to guarantee independence. Datasets must be different for each model, which will give us more precise results while aggregating each model. Also, by using ensemble techniques bias and variance can be reduced in standard learning algorithms.

A block diagram to show how ensemble models work is given in Fig-2, where the same or different subsets of the training datasets are fit into different models and their corresponding prediction is further aggregated for final output. In Fig-3, we can see the most commonly used ensemble learning techniques and

different algorithms in it. In generally commonly used ensemble techniques are bagging, boosting, and stacking.

- **Bagging**: Bagging (also called bootstrap aggregation) is a significant, effective, and simple ensemble strategy. It uses numerous versions of a training set by using the bootstrap that means sampling with replacement. It is also seen that the bagging technique is more effective when it is applied to non-linear models. In generally Bagging is used with a Decision tree classifier to reduce variance. In this approach, several subsets of the training sample are randomly selected and each subset of the training data is used for training the decision trees. This generates an ensemble of various models and finally, the outcome of every model is aggregated with the help of majority voting or averaging for a high-level accuracy. Different research works have proven that the bagging technique can significantly increase the performance of a weak classifier. Bagging can be used with any category of model for classification or regression.

Table-1: Different research papers and their activities

Paper Number	Aim of the research	Dataset used	Algorithms applied	Accuracy level in %	Remarks
Mamta Sharma [13]	Prediction of Heart disease	N/A	Naïve Bayes	90.74	Neural Network performed best with 15 attributes
			Decision Tree	99.62	
			Neural Network	100	
Thirumal P. C. and Nagarajan N [14]	Diagnosis of diabetes	PIMA Indian diabetes dataset	Naïve Bayes	77.86	C.45 algorithm performed best but others are also good in comparison with each other
			SVM	77.47	
			C4.5	78.25	
			k-NN	77.73	
Jyoti Soni and more [15]	Prediction of heart disease	Cleveland Heart Disease database	Naïve Bayes	86.53	Decision tree obtained the best results among all. Bayesian classifier also obtained similar prediction accuracy like decision tree.
			Decision Tree	89	
			Artificial Neural Network	85.53	
K. Manimekalai [16]	Prediction of heart disease	N/A	Naïve Bayes	99.52	Naïve Bayes algorithm when applied with WEKA 3.6.4 gives the highest accuracy in comparison with others. Also found that WEKA tool is very easy to implement with the best performance
			J48 (Decision Tree)	95.56	
			Neural Network	79.19	
			Fuzzy Logic	83.85	
			Decision Tree	52.33	
			Neural Network	96.5	
Priya B. Patel [1]	Prediction of diabetes	Pima Indian Diabetic Set	Gaussian Naive Bayes	69.685	Pima Indian diabetic dataset from the UCI is used and the result showed that the K-NN algorithm performed best
			K-NN	70.866	
			SVM	64.137	
			Decision Tree	67.176	
Amit Tate et al [5]	Prediction of disease	Demo dataset	Naïve Bayes	N/A	Random forest algorithm performed best concerning all factors
			Random forest (Ensemble Classifier)		
			SVM		
Durga Kinge et al [8]	Prediction of disease	UCI Heart-disease dataset	J48	78.15	Their result showed that Random forest, Naïve Bayes, and simple logistic performed very well in the prediction of heart disease.
			Naïve Bayes	82.59	
			Simple Logistic	83.1	
			Multilayer Perceptron (MLP)	79.41	
			Bagging	81.59	
			AdaBoost	81.59	
			Random Forest	83.15	
M.A.Nishara Banu et al	Prediction of Heart disease	Heart disease dataset with	K-mean based MAFIA	74	K-mean based MAFIA (Maximal Frequent Itemset

[10]	using data mining Classification Algorithm	19 attributes	K-mean based MAFIA with ID3	85	Algorithm) with ID3 and C4.5 performed best in comparison with others
			K-mean based MAFIA with ID3 and C4.5	94	
C. Beulah et al [20]	Prediction of Heart disease risk	Cleveland Heart Disease database	Ensemble Classifier	85.48	The performance of Majority vote algorithm RF, NB, Bayesian Network, and Multilayer Perceptron was best.
Ibomoiye Domor Mienye et al [21]	Prediction of Heart disease risk	Cleveland and Framingham heart disease dataset	Ensemble Classifier	93	Randomized decision tree ensemble performed best
Anurag Kumar Vermaa et al [22]	Skin disease prediction	UCI Skin disease dataset	Ensemble Classifier	99.68	Gradient Boosting ensemble method applied on Radius Neighbors Classifier performed best with Feature Selection

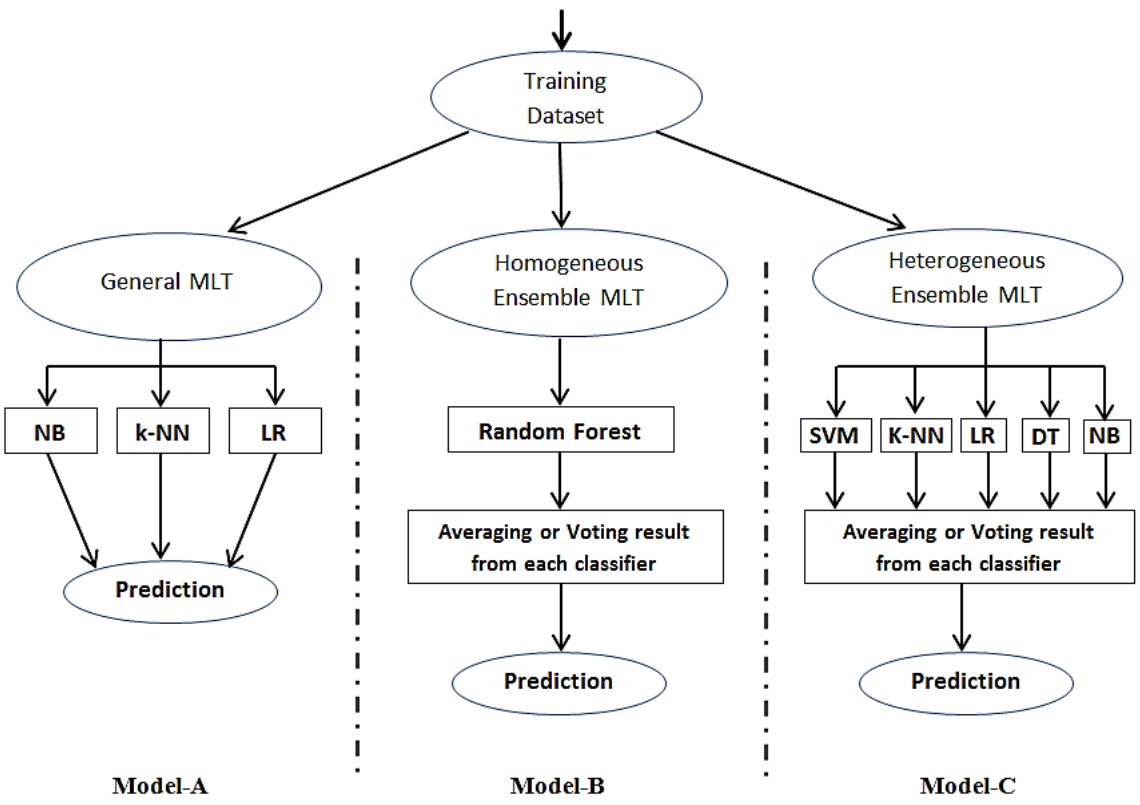

Fig-1: Proposed disease prediction approach with three models

Volume 10, Issue 4
August 2021

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

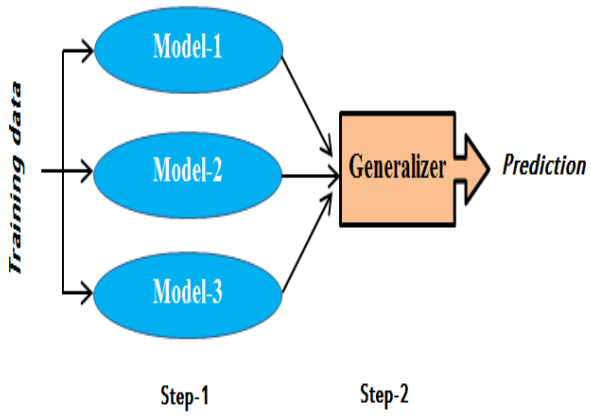

Fig-2: Basic working principle of the ensemble model

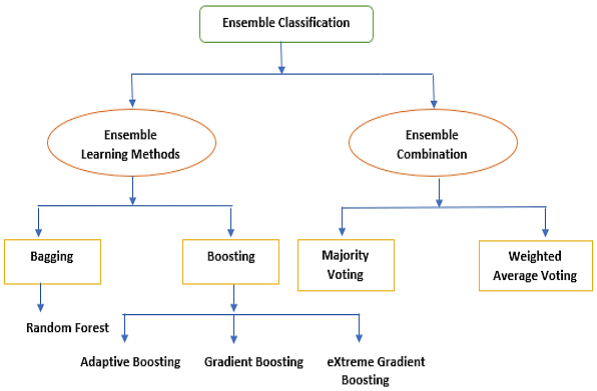

Fig-3: Commonly used ensemble techniques [19]

- **Boosting:** It is a sequential ensemble method and is the most frequently used technique with the most powerful learning ideas. In the sequential model application technique**,** the set of combined weak models are not directly fitted independently from each other. Here the models are fit iteratively in such a way that the training of a model at a particular step depends on the models fitted previously. Though this method is primarily designed for classification problems it is also found useful in regression. Boosting is an iterative technique that adjusts the weight of observation as per the last classification. When an observation is not correctly classified, boosting tries to increase the weight of that observation. Its main idea is to minimize the bias error and convert weak learners to a strong predictive model.

- **Stacking:** It is another important and widely used ensemble machine learning technique. In this approach, two important phases are there. In the first phase, several models are trained with different algorithms types on a single training dataset. Again, in the second phase, instead of using the single best model from the outputs, all models' outputs are aggregated using another model which is also called Meta learner to get the final prediction. Here the base-learners outputs become inputs for the Meta learner.

5. EXPERIMENTAL APPROACH

In this study, we have proposed three experimental models (A, B, and C) through which the heart disease dataset is tested and checked whether the ensemble technique performs better or not. For this experiment, the Cleveland heart dataset from the UCI machine learning repository is used and is downloaded from the Kaggle website. Originally it contains 76 attributes and a total number of 303 instances, but it is found that only 14 features are important for research. The details of the attributes are mentioned in table-2. The age group of patients range from 29 to 79 and the final result of this dataset i.e. whether a patient is having heart disease or not is indicated by the attribute Num. Its value is binary (0 and 1), 0 indicates no heart disease and 1 indicates the patient is having heart disease. Here out of 303 patient data 46% are having problems with heart disease and the remaining 54% are not. In Fig-4, a bar diagram is shown to represent the target value concerning the male and female patient data in our dataset. For training and testing purposes we have used 75% of the total instances as a training dataset and another 25% as testing data. The same training dataset is fitted to train all three models.

In our experiment, the Python machine learning tool is used for the implementation and analysis of the result. The basic idea behind this analysis is to compare the prediction accuracy of the three proposed techniques and to compare the outcomes to check whether the ensemble technique can improve the weak classifiers' performance or not. The following figure (Fig-5) shows the inter-relation among the attributes, which indicates how attributes are correlated to each other, either negatively or positively. Similarly, we can also represent the attributes graphically with respect to their values as shown in the Fig-6.

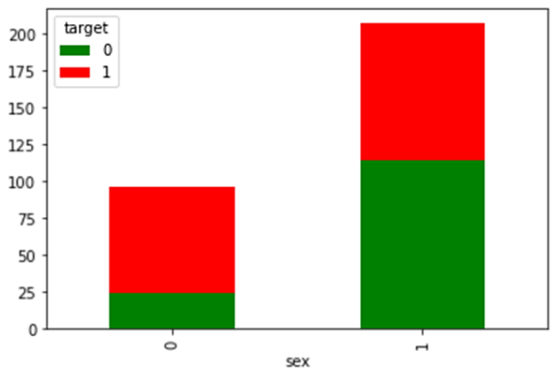

Fig 4: Target value with respect to Sex attribute of the dataset

Volume 11, Issue 1
February 2022

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Table-2: Different features of the UCI Cleveland heart disease dataset and its description

Attribute No.	Name of Attribute	Description
1	age	Patient age in years
2	gender	Gender of Patient (1-Male, 0-Female)
3	cp	Chest pain type: 0-Typical angina, 1-Atypical angina 2-Non-anginal pain, 3-Asymptomatic
4	trestbps	Resting blood pressure (in mm Hg)
5	chol	Serum cholesterol (in mg/dl)
6	fbs	Fasting blood sugar is >120 (in mg/dl) 1-true, 0-false
7	restecg	Resting electrocardiographic results 0: Nothing to note, 1: ST-T Wave abnormality 2: Possible or definite left ventricular hypertrophy
8	thalash	Maximum heart rate achieved
9	exang	Exercise-induced angina (1-Yes, 0-No)
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	Slope of the peak exercise ST segment 0-Upsloping (better heart rate with exercise) 1-Flatsloping (typical healthy heart) 2-Downslopins (signs of the unhealthy heart)
12	ca	Number of major vessels (0-3) colored by flourosopy
13	thal	Thalium stress result 3-Normal, 6-Fixed defect, 7-Reversible defect
14	num (Target Attribute)	The patient is having heart disease or not 0-No, 1-Yes

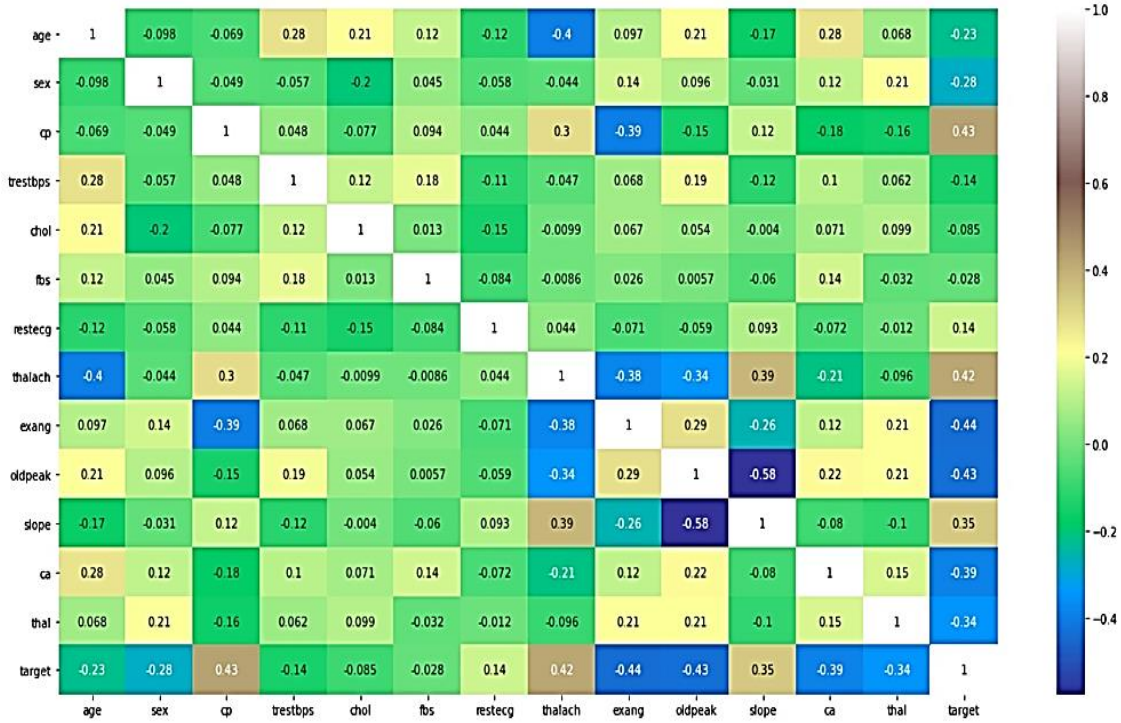

Fig 5: Correlation among the attributes

Volume 10, Issue 4
August 2021

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

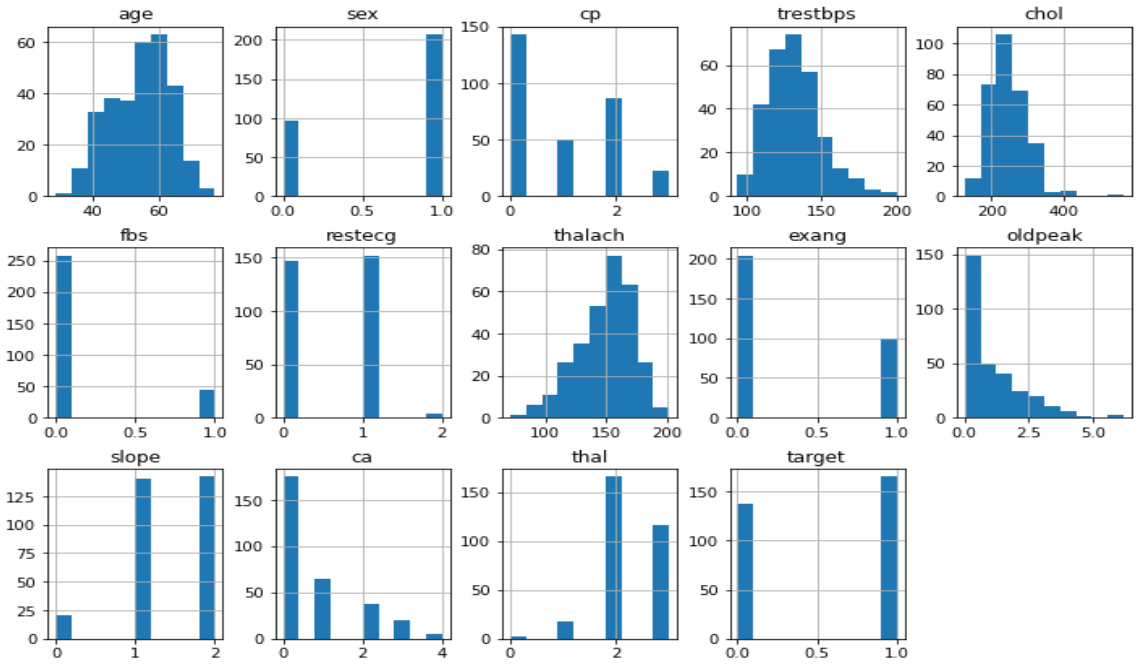

Fig-6: Graphical representation of each attribute

In the first model (Model-A), three well-known classifiers Naïve Bayes, k-nearest neighbor, and Logistic regression are used to check their accuracy level for heart disease prediction. As this model is very simple and straight so it is referred to here as general MLT. The overall comparison of all the three classifiers is shown in Table-3. The k-NN classifier scored best when the k value is 7. After implementation, the Logistic Regression classifier produces a maximum of 87.91% accuracy among all of them.

Table-3: Comparison of classifiers of model-A

Algorithm	Accuracy Score	Target value	Precision	Recall	F1-score	Support
NB	86%	0	87	80	84	41
		1	85	90	87	50
LR	87.91%	0	94	78	85	41
		1	84	96	90	50
K-NN	73.62%	0	76	61	68	36
		1	70	82	76	40

In Model-B, a Random forest (RF) ensemble classifier is used. It is also found from the different research activities that RF classifier is very efficient in disease prediction. The output of model-B is shown in Table-4.

Finally, in the third model (Model-C) the ensemble of five different classifiers is used to achieve an improved output, and this technique is referred to as heterogeneous ensemble MLT. In the experiments with models B and C, Bagging and Majority voting ensemble algorithms are applied. Majority voting is one of the most important ensemble strategies which combine the output of multiple classifiers to get the proper output result. The following table shows the result of model C.
After successful implementation of all the three models with the Cleveland heart disease dataset, it is seen that model-3 provides the highest accuracy for the detection of heart disease. Here we have used five different base classifiers and when their outputs are combined with the majority voting technique the outcome gives us a maximum accuracy score of 92.1%. This indicates that the ensemble technique improves the accuracy of weak classifiers. The bar diagram (Fig-7) shows the comparison of our proposed model's performance.

Table-4: Analysis of Model-B approach

Algorithm	Accuracy Score	Target value	Precision	Recall	F1-score	Support
RF	89.47%	0	94	83	88	36
		1	86	95	90	40

Table-5: Analysis of Model-C approach

Algorithm	Accuracy Score	Target value	Precision	Recall	F1-score	Support
Ensemble of DT, k-NN, NB, LR, SVM	92.10%	0	97	86	91	36
		1	89	97	93	40

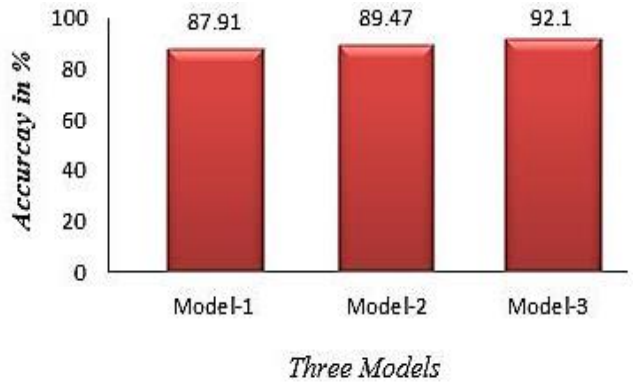

Fig-7: Accuracy scores of the three models

6. CONCLUSION

Comparison of algorithms is always a critical task as we can see their behavior changes with different situations. Several research activities, mostly related to medical datasets, have been analyzed to find the best algorithms in terms of efficiency and performance. In most of the cases we have seen Python and Weka tool is primarily used to analyze the datasets for the prediction of a particular disease. This paper initially compares the different research activities related to techniques and methods used by the researchers. It is seen that in most of the activities where ensemble techniques are used, improves the performance of the classifier. For this purpose, this study also performs three different techniques (Model-A, Model-B, and Model-C) to check whether the ensemble technique improves the accuracy level of weak classifiers or not. After implementation, it is seen that model-3, which is an ensemble of five classifiers gives the maximum accuracy of 92.1%. It is also important to consider that algorithms can perform best only when the data selection is proper; that means we have to collect the proper dataset with a minimum number of attributes that are most relevant to the target value. In our future study, an improved feature selection technique will be used to increase the classifier performance for the prediction of cardiovascular disease.

ACKNOWLEDGEMENT

The authors recognize the huge assistance got from the researchers whose articles are referred to and remembered for references of this manuscript. Also, the authors are thankful to writers/editors/distributors of those articles, diaries, and books from where the writing for this article has been discussed as well as reviewed. The primary author likewise prefers to express gratitude toward Assam down town University for giving us a better chance in this study.

REFERENCES

[1] Priya B. Patel, Parth P. Shah, Himanshu D. Patel, Analyze Data Mining Algorithms for Prediction Of Diabetes, International Journal of Engineering Development and Research, 2017, Volume 5, Issue 3, ISSN: 2321-9939.

[2] V. Kirubha, S. Manju Priya, Survey on Data Mining Algorithms in Disease Prediction, International Journal of Computer Trends and Technology (IJCTT), August 2016, Volume-38, Number 3.

[3] Priyanka P. Shinde, Kavita S. Oza, Rajanish K. Kamat, An Analysis of Data Mining Techniques in Aggregation with Real-Time Dataset for the Prediction of Heart Disease, International Science Press, 2016, I J C T A, 9(20), pp. 327-336.

[4] Meenu Singla, Kawaljeet Singh, Heart Disease Prediction System using Data Mining Clustering Techniques, International Conference on Advances in Emerging Technology (ICAET 2016), International Journal of Computer Applications (0975 – 8887).

[5] Amit Tate, Bajrangsingh Rajpurohit, Jayanand Pawar, Ujwala Gavhane, Comparative Analysis of Classification Algorithms Used for Disease Prediction in Data Mining, International Journal of Engineering and Techniques, Nov-Dec-2016, Volume 2 Issue 6.

[6] Latha Parthiban and R. Subramanian, Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm, International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering, 2007, Vol.1, No.5.

[7] Lazim Abdullah, A fuzzy linear regression model for identifying risk factors of coronary heart disease, World Journal of Modelling and Simulation, Vol. 13 (2017) No. 2, pp. 83-93.

[8] Durga Kinge, S. K. Gaikwad, Survey on data mining techniques for disease prediction, International Research Journal of Engineering and Technology (IRJET), Jan-2018, Volume: 05 Issue: 01, e-ISSN: 2395-0056.

[9] K. Aparna, Dr. N. Chandra Sekhar Reddy, I. Surya Prabha, Dr. K. Venkata Srinivas, Disease Prediction in Data Mining Techniques, International Journal of Computer Science and Technology, April - June 2014, Vol. 5, Issue 2, ISSN: 0976-8491 (Online), 2229-4333 (Print).

[10] M.A. Nishara Banu, B Gomathy, Disease Predicting System Using Data Mining Techniques, International Journal of Technical Research and Applications, Nov-Dec 2013, e-ISSN: 2320-8163, Volume 1, Issue 5, PP. 41-45.

[11] Saurabh Pal, Vikas Chaurasia, Is Alcohol Affect Higher Education Students Performance: Searching and predicting pattern using Data Mining Algorithms, International Journal of Innovations & Advancement in Computer Science, ISSN 2347 – 8616, Volume 6, Issue 4, April 2017.

[12] Akanksha.A. Pande, S. A. Kinariwala, Analysis of Student Learning Experience by Mining Social Media Data, International Journal of Engineering Science and Computing, May 2017, Volume 7 Issue No.5.

[13] Mamta Sharma, Farheen Khan, Vishnupriya Ravichandran, Comparing Data Mining Techniques Used for Heart Disease Prediction, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056, Volume: 04 Issue: 06, June -2017.

[14] Thirumal P. C. and Nagarajan N, Utilization Of Data Mining Techniques For Diagnosis Of Diabetes Mellitus - A Case Study, ARPN Journal of Engineering and Applied Sciences, Vol. 10, No. 1, January 2015, ISSN 1819-6608.

[15] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni, Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction, International Journal of Computer Applications (0975 – 8887), Volume 17– No.8, March 2011.

[16] K. Manimekalai, A Proficient Heart Disease Prediction Method Using Different Data Mining Tools, International Journal of Engineering Science and Computing, March 2016, ISSN 2321 3361, Volume 6 Issue No. 3.

17] https://www.usfhealthonline.com/resources/key-concepts/data-mining-in-healthcare/

[18] http://www.justscience.in/articles/can-data-mining-healthcare-industry-make-us-ealthier

[19]https://medium.com/analytics-vidhya/credit-card-fraud-detection-with-bagging-ensemble-learning-cc337f3d8aa2 27/05/2021

[20] C. Beulah Christalin Latha, S. Carolin Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, Informatics in Medicine Unlocked 16 (2019) 100203, Published by Elsevier Ltd., https://doi.org/10.1016/j.imu.2019.100203.

[21] Ibomoiye Domor Mienye, Yanxia Sun, Zenghui Wang, An improved ensemble learning approach for the prediction of heart disease risk, Informatics in Medicine Unlocked 20 (2020) 100402, Published by Elsevier Ltd., https://doi.org/10.1016/j.imu.2020.100402.

[22] Anurag Kumar Vermaa, Saurabh Palb, Surjeet Kumarb, Comparison of skin disease prediction by feature selection using ensemble data mining techniques, Informatics in Medicine Unlocked 16 (2019) 100202, Published by Elsevier Ltd., https://doi.org/10.1016/j.imu.2019.100202

AUTHOR PROFILES

Mr. Bhabesh Deka completed Master of Computer Application (MCA) from Dibrugarh University, Assam, India. He is presently a research scholar of Assam down town University, India and also working as a lecturer of Computer Science and Application in M.N.C. Balika Mahavidyalaya, Nalbari, Assam. His area of interest includes Machine learning, Data mining, and Data Science.

Dr. Bismita Choudhury, presently working as an Assistant Professor of Computer Science and Engineering in Assam down town University, India. She has completed B.Tech (CSE) from North Eastern Hill University, Shillong in the year 2011 and M.Tech (CSE) from Assam Don Bosco University (2014). After then she has completed her Ph.D. from Swinburne University of Technology, Australia in the year 2019. Her area of interest includes AI, Deep Learning, Medical Image processing, Biometrics, Steganography, and Steganalysis.