Premature Detection of Cardiomegaly using Hybrid Machine Learning Technique

Bhanu Prakash Doppala, Department of Computer Science and Multimedia, Lincoln University College, Malaysia,

Midhunchakkravarthy, Department of Computer Science and Multimedia, Lincoln University College, Malaysia,

Debnath Bhattacharyya, Department of Computer Science and Engineering, K L Deemed to be University, Koneru Lakshmaiah Education Foundation, Guntur, India.

Abstract-The clinical field usually handles large measures of information. Taking care of tremendous information by conventional techniques can influence the outcomes. Utilization of calculations for artificial intelligence to discover realities in clinical research, mainly for the prediction of a particular disease. The early acknowledgement of the infection is vital for the examination of patient meds and experts. Utilizing machine learning techniques can prompt a quick ailment prediction system with high accuracy. In the medical area, machine learning applications playing a crucial role in predicting diseases. This particular paper assesses different classifiers used for the expectation of cardiovascular infirmities. There are major machine algorithms; for instance, Decision Tree, Random Forest, is used for envisaging heart diseases. We presented a forecast model with various features with different combinations and a few known grouping strategies. We produced an upgraded performance level with an accuracy of 84:42%.

Keywords-Heart Disease, Machine Learning, Cardiomegaly, CVD, ensemble learning; Prediction.

Introduction

Machine Learning is an artificial intelligence branch that aims to provide computer methods to accumulate change and update system knowledge. Artificial intelligence (A.I.) allows systems to observe from environments, execute certain features, and increase the likelihood of success in fixing real-world challenges. A.I. turns out to be an exciting field with technological improvements and scientific growth. It, therefore, leads to growing attention on ML techniques.

The forecast for diseases plays a vital role in machine learning. Prediction of various types of conditions using ML techniques, and here we look at how machine learning techniques used to foresee different sickness types. This paper focused on the prediction of heart disease.

There are different scopes of heart sicknesses separated from a coronary episode, which is aggregately called cardiovascular maladies. There are numerous purposes behind the improvement of heart sicknesses, for example, smoking, glucose, obesity, and sadness, elevated cholesterol, less than stellar eating routine, and hereditarily relative. There are numerous kinds of heart sicknesses, for example, angina, arrhythmia, intrinsic coronary illness, fibrillation, coronary supply route infection, cardiovascular breakdown, and fibrillation. [10] The desire for heart diseases causes us to treat the patient before the patient shows up at a cardiovascular breakdown.

Cardiac disease is one of the most common illnesses which may result in impairment, death, & other economic crises in patients who are afflicted by cardiovascular disease. To substantiate the reality, and there was a report published by WHO that quotes 'Annually, 17.5 million individuals are dying out as a result of CAD (Coronary Artery Disease) in this world'. So the available amount of information regarding these diseases in the database has become exceedingly dense and challenging to analyze by the healthcare centres. So with the help of swiftly growing technologies and making use of Machine learning algorithms, the job of processing different details in healthcare (hospitals along with other medical) centres can be eased to a great extent.

Numerous individuals lost their lives because of heart infections. Constant cardiovascular breakdown (CHF) is regular in innumerable individuals. Lately, to distinguish coronary illness, precisely clinical gear is created. Computer-based intelligence techniques find a response to this issue by perceiving coronary ailment accurately.

Literature Survey

Qualitative research took place on the coronary ailment dataset. Most of the models turned with better accuracy towards the prediction of a heart ailment.

DOI: 10.5373/JARDCS/V12I6/S20201054
**Corresponding Author: Debnath Bhattacharyya, Email: debnathb@kluniversity.in*
Article History: Received: Mar 08, 2020, Accepted: June 10, 2020

Vincy Cherian et al. proposed a technique where the model is accepted to fend off unessential indicative tests led and deferral in beginning the correct treatment. [1] Therefore, by early diagnosing patients with cardiovascular disease, can be saved with time and money. In this way, specialists can analyze without pointless treatment incited because of the inability and instincts made by a specialist. The system can reach out to fuse different regions for different affliction desires.

Sushmita Manikandan et al. proposed a coronary failure expectation framework utilizing the heart disease archive of UCI to think of a model. They have created a model having a place with the structure which orders the hazard factor in an individual. [2]Using different tools of data mining prediction and analysis can be done. A model built to assess the connection between the covered examples just as will, in general, make a forecast.

Creators had exhibited the attainability study and the advancement of coronary illness grouping installed framework. The collection signifies common heart illnesses, for example, angina, dead myocardial tissue, and coronary corridor ailments. The framework could likewise be utilized unexpected "on obligation" doctors, of any specialized topic, and could manage the cost of the first, or beginning analysis of any cardio-apathy. [6] On the off chance that any framework distinguishes a heart issue, this framework enriches with better sickness conclusion PPV assessed to different declarations, and this way, it tenders raised affirmation than different strategies. Another chief consideration is the truth that this framework was practically equivalent to numerous different contexts by getting into the full informational index, and this framework practised a fluffy grouping calculation to reduce the informativelist, consequently alleviating its utilization.

This paper [11] proposes that enormous information existing from clinical diagnosis using data mining tools, and pulls out significant data known as information. Mining is a strategy for reviewing large sets of information to obtain the examples which are covered up and in the past obscure affiliations and information discovery to encourage the improved comprehension of clinical data to obstruct coronary illness. This paper offers a quick and basic assessment and view of reachable prediction models utilizing information mining from 2004 to 2016.

In research [3], the author modified the SVM classification based on R.S. for the prediction of medical diseases. The approach used the benefits of R.S. to eliminate redundant information and the advantage of SVM in training and testing data. This technique used on three sets of data, which shows an increase in accuracy as compared to other approaches.

Problem Statement

In Prior research, they have broken down the utilization of Machine learning methodologies for estimating and classifying the heart illness. Likewise, hardly any analysts endeavour to utilize hybrid optimization techniques for upgraded classifications in machine learning. The most prospective studies in the writing misuse streamlined procedures, for example, Particle Swarm Optimization and Ant Colony Optimization with a particular ML strategy, for example, SVM, KNN, or Random Forest.

Machine Learning Techniques Inhealthcare

Decision Tree: This Structure will have root and leaf nodes. Each interior node used to mean a trial of a trait, though the branch indicates the result of the test and leaf level. The root hub represents the highest point in the tree. Among many structures available J48 is well known. [12].

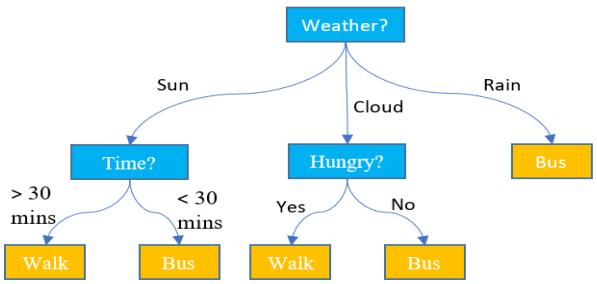

Fig. 1 Decision Tree Representation[20].

Random Forests: are a group learning strategy for classification and regression procedures. It generates several Decision trees during the training time and produces the class that is the method of the types yield by individual trees. It likewise tries to reduce the issues of high bias and high variance by averaging to locate a characteristic

harmony between the two extremes. We can quickly implement this technique on R and Python with the help of inbuilt packages.

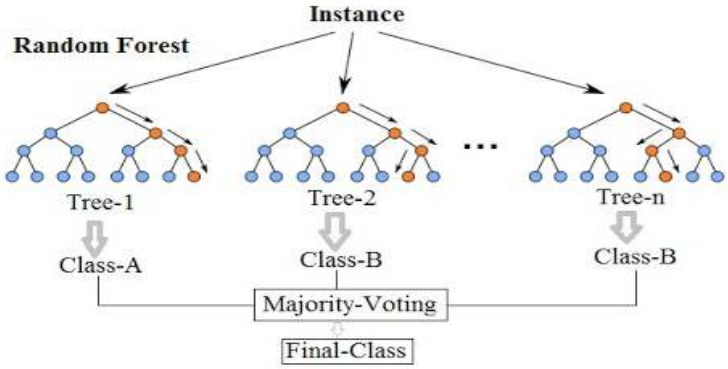

Fig. 2 Random forest classifier[21].

Logistic Regression

It is a measurable thought used to choose the weight of the relationship between one ward variable (generally implied by Y) and a movement of other advancing factors (known as independent components). Two fundamental sorts of backsliding are linear and multiple linear regression. In like manner, there are a couple of non-linear regression techniques used for logically befuddled data assessment.

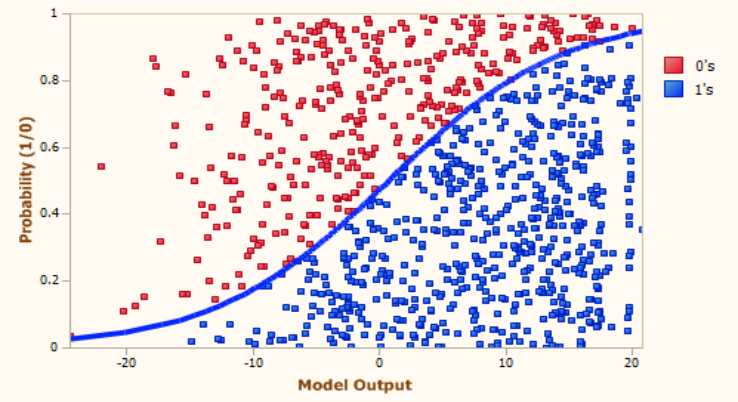

Fig. 3 Logistic Regression[22].

Ensemble D.M. approach

For progressive and exact forecast results, an outfit technique is a very much demonstrated methodology to accomplish profoundly precise classification of information by hybridizing various classification mechanisms. The enhanced performance is an essential in-manufactured component of the ensemble technique.

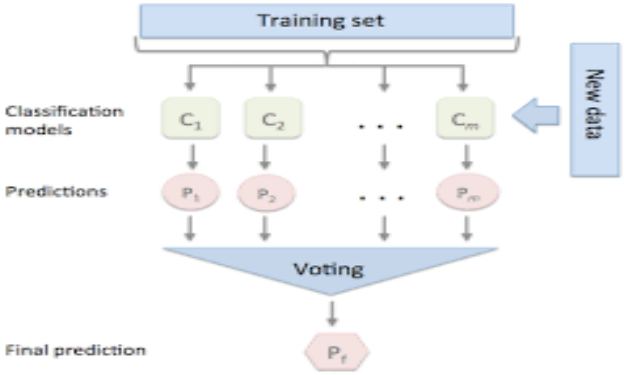

Fig. 4 Ensembles classification[23].

This following work suggests a weighted vote-based classifier ensemble system, and get the better constraints of customary D.M. strategies by utilizing the ensemble of three heterogeneous classifiers: Decision Tree, Gaussian

DOI: 10.5373/JARDCS/V12I6/S20201054
*Corresponding Author: Debnath Bhattacharyya, Email: debnathb@kluniversity.in
Article History: Received: Mar 08, 2020, Accepted: June 10, 2020

Naive Bayesian, and SVC. We have used the coronary illness dataset taken from the UCI archive from Cleveland repository.

Proposed System

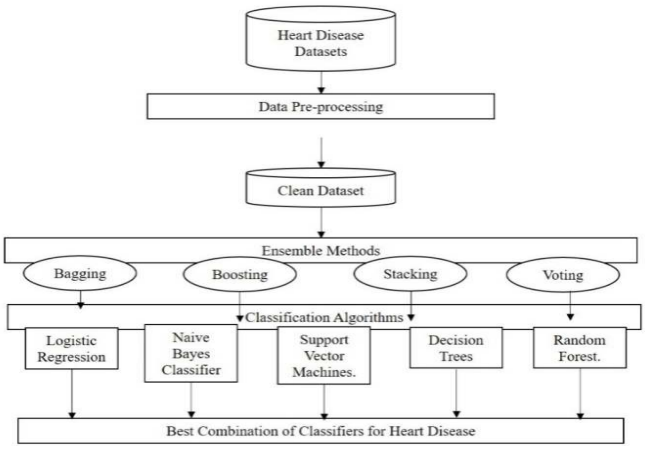

Fig. 5 Proposed system architecture

Dataset Description

We have taken Heart disease UCI dataset from Kaggle for this study paper. This catalogue encloses 14 characteristics.

Dataset obtained from the UCI Repository [3], which has 303 records with fourteen attributes. We used the WEKA tool for data analysis of this data set. Here attribute num is the class attribute that is used by this prediction system to identify heart disease.

Table. 1 Heart disease dataset[19].

S.No	Name	Description
1	age	age in years (28 - 77)
2	sex	0- female , 1- male
3	cp	chest pain type 1-typical angina, 2-atypical angina, 3-non-angina pain, 4-asymptomatic
4	trestbps	resting blood pressure (mm/Hg) (0 - 200)
5	chol	serum cholesterol (mg/dl) (0 - 603)
6	fbs	fasting blood sugar 0-false(< 120 mg/dl) , 1-true (> 120 mg/dl)
7	restecg	resting electrocardiographic results 0-normal , 1- having ST-T wave abnormality 2- showing probable or definite left ventricular hypertrophy by Estes' criteria
8	thalach	maximum heart rate achieved (60 - 202)
9	exang	exercise induced angina 0-no 1-yes
10	oldpeak	ST depression induced by exercise relative to rest (mm) (-2.6 – 6.2)
11	slope	the slope of the peak exercise ST segment 1-upsloping , 2-flat , 3-downsloping
12	ca	number of major vessels colored by fluoroscopy (0-3)
13	thal	3-normal 6-fixed defect 7-reversable defect
14	num	diagnosis of heart disease (angiographic disease status) – target attribute 1. - negative diagnosis (absence) 2. - 4 (from least serious most serious - presence)

Information pre-processing

Deals with cleaning the dataset, missing qualities, and prohibit exceptions. It began with transferring the dataset into the design segment utilizing the "Retrieve" operator followed by the utilization of "Set Role" operator for determining the class label lastly the "Replace Missing Values" by replacing them with the mean of different values present for that specific component in the dataset.

Classification

It is simpler to recognize the correct class, and the outcome would be more precise than with the clustering method. With the end goal of the comparative investigation, we used three Machine Learning calculations. The diverse Machine Learning (ML) algorithms are Decision Tree, Gaussian Naive Bayesian, and SVC. The motivation to pick these calculations depends on their popularity [15].

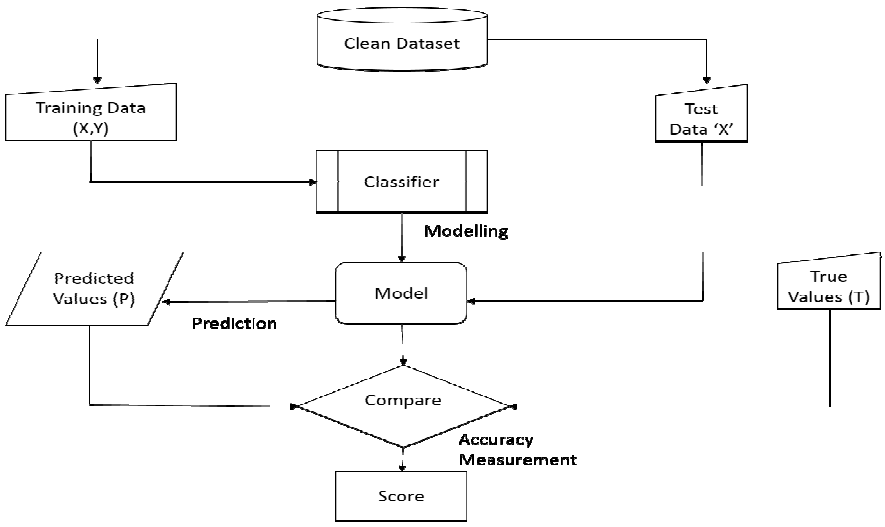

Fig. 6 Process of data classification and model accuracy computation of general classifier.

Result & Analysis

In this exploration work, determination of coronary illness test acquired from the UCI information vault. The dataset set comprises of fourteen (14) features which contain 303 samples. The dataset partitioned into a preparation set and testing set. The dataset isolated utilizing proportion 60:40, for example, 60% of the dataset for preparing and 40% of the dataset for testing of the system, which is the standard proportion for partitioning dataset in the A.I. dataset. Here Jupiter notebook has been utilized for exploratory investigation of CVD informational indexes and execution assessment of every classifier in which applied for the arrangement of informational indexes

The following illustration represents the dataset after loading it for the necessary operations.

```
In [ ]: df = pd.read_csv('heart.csv')

In [4]: df.head()

Out[4]:
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Fig. 7 Dataset Description

Figure 7. Shows the description of the features and their values after loading the dataset into Jupiter Notebook. A total number of 14 features and 303 samples had been packed, out of which displayed the top 5 values. Here we represented the Detail distribution of the dataset attributes in Figure 8.

DOI: 10.5373/JARDCS/V12I6/S20201054
*Corresponding Author: Debnath Bhattacharyya, Email: debnathb@kluniversity.in
Article History: Received: Mar 08, 2020, Accepted: June 10, 2020

```
In [8]: plt.figure(figsize=(14,8))
        sns.heatmap(df.corr(), annot = True, cmap='coolwarm',linewidths=.1)
        plt.show()
```

Fig. 8 Detail distribution of the dataset attributes

Confusion Matrix:

Confusion Matrixrepresents the performance of a classifier.

Table. 2 Representation of Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives(TPs)	False Positives(FPs)
Predicted Negative (0)	False Negatives(FNs)	True Negatives(TNs)

With the assistance of the confusion matrix, the simultaneous measurements are determined.

$$Precision = \frac{True\ Positive}{Total\ Predicted Positive} \quad (1)$$

$$Recall = \frac{True\ Positive}{Total\ Actual\ Positive} \quad (2)$$

$$F - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

$$Specificity = \frac{True\ Negative}{Total\ Actual\ Negative} \quad (4)$$

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total} \quad (5)$$

With the help of a confusion matrix for individual models, we derived the following values.

Table. 3 Result Comparison between Proposed and Existing Classification Techniques

Name of the Model	Recall	Specificity	Precision	F-Score
Decision Tree	0.77	0.78	0.74	0.78
Random Forest	0.76	0.81	0.79	0.77
Logistic Regression	0.87	0.79	0.72	0.78
Gradient Boosting	0.9	0.76	0.66	0.77
Extreme Gradient Boosting	0.85	0.76	0.68	0.75
Proposed Model	0.88	0.82	0.77	0.82

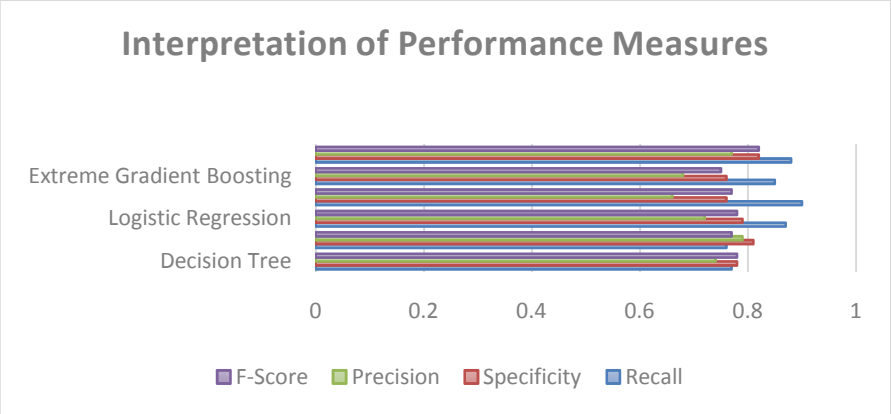

Fig. 9 Graphical Representation of Accuracy between Existing and Proposed Techniques

The table below exhibits the performance differentiation of the classification techniques used.The accuracy of the proposed system is 84.42% for 14 characteristicsout of all algorithms proposed system performed better in terms of accuracy.

Table. 4 Model Comparison for Accuracy

Name of the Model	Accuracy
Decision Tree	77.86
Random Forest	81.15
Logistic Regression	81.96
Gradient Boosting	80.14
Extreme Gradient Boosting	79.50
Proposed Model	84.42

DOI: 10.5373/JARDCS/V12I6/S20201054
*Corresponding Author: Debnath Bhattacharyya, Email: debnathb@kluniversity.in
Article History: Received: Mar 08, 2020, Accepted: June 10, 2020

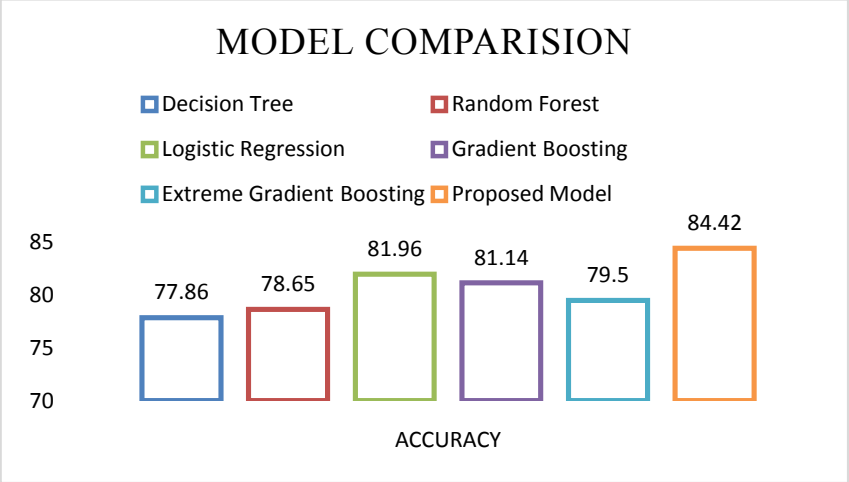

Fig. 10 Accuracy Comparison between models

Finally, in the table above, it is shown the Interpretation of Performance Measures with the help of the Confusion Matrix [16, 18]. We identified an increase in accuracy with the proposed system compared to the existing machine learning techniques.

We analyzed the result with the help of the ROC curve for better understanding.

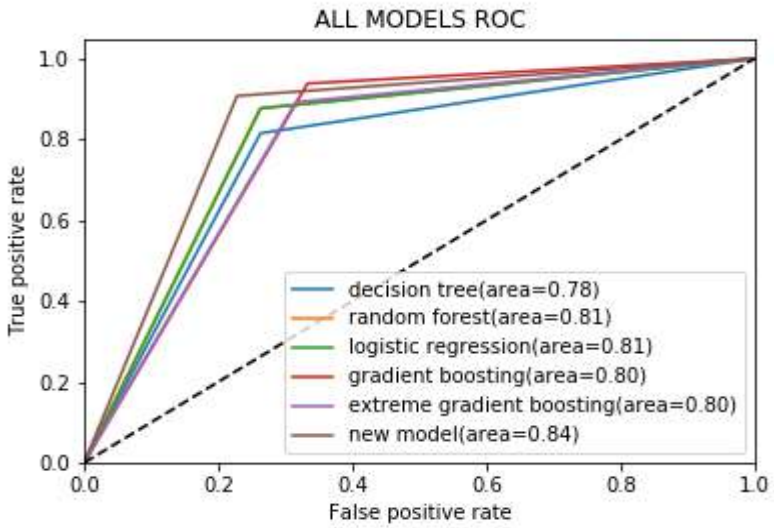

Fig. 11 Roc curve for the models.

Conclusion

We proposed a model to identifycoronary illness with greater exactness. Efficiency has been enhanced to 84.42 % by the proposed model, which has given increased results over other machine learning algorithms. Still, there are improvements which can be explored towards the improvement of accuracy.Asfuture work, we can utilize this framework for the examination of various informational collections. Heart diagnosiscan be improved fundamentally by taking care of countless class labels during the process of prognosis, and it tends to be another positive course of research.

References

[1] V. Cherian and B. M.S., "Heart Disease Prediction using Naive Bayes Algorithm & Laplace Smoothing technique," International Journal of Computer Science Trends & Technology, vol. II, no. 2, pp. 68-73, 2017.

[2] S. Manikandan, "Heart Attack Prediction System," International Conference on Energy, Communication, Data Analytics & Soft computing, pp. 817-820, 2017.

[3] De Carvalho Junior, Helton Hugo, et al. "A heart disease recognition embedded system with fuzzy cluster algorithm." Computer methods and programs in biomedicine 110.3 (2013): 447-454.
[4] Wghmode, Mr Amol A., Mr Darpan Sawant, and Deven D. Ketkar. "Heart Disease Prediction Using Data mining Techniques." Heart Disease (2017).
[5] Cp, Prathibhamol, Anjana Suresh, and Gopika Suresh. "Prediction of cardiac arrhythmia type using clustering and regression approach (P-CA-CRA)." Advances in Computing, Communications, and Informatics (ICACCI), 2017 International Conference on. IEEE, 2017.
[6] Banu, N.K. Salma and Suma Swamy. "Prediction of heart disease at an early stage using data mining and big data analytics: A survey." Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 016 International Conference on. IEEE, 2016.
[7] Zhang G., "A Modified SVM Classifier Based on R.S. in Medical Disease Prediction," 2009.
[8] M. Sultana, A. Haider, and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction," 2016 3rd Int. Conf. Electr. Eng. Inf. Commun. Technol. IEEE ICT, 2016, 2017.
[9] Bhanu Prakash Doppala, Divya Midhunchakkaravarthy, Debnath Bhattacharyya, "Early Stage Detection of Cardiomegaly: An Extensive Review, "IJAST, Vol 125, pp 13-24.
[10] Divya Annepu, Gowtham G, "Cardiovascular Disease Prediction Using Machine Learning Techniques," IRJET, Volume: 06 Issue: 04.2019.
[11] Mr Chala Beyene, Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", IJPAM, Volume 118 No. 8 2018, 165-174.
[12] Sanchita Chatterjee, Yasha Jaggi, B.Sowmiya, "Survey on Prediction of Heart Disease Using Data Mining," ICISS-21-22 Feb. 2019
[13] Mustafa Jan, Akber A Awan, Muhammad S Khalid, Salman Nisar, "Ensemble approach for developing a smart heart disease prediction system using classification algorithms"Research Reports in Clinical Cardiology, Jan 2019.
[14] Vikas Chaurasia and Saurabh Pal "Data Mining Approach toDetect Heart Diseases", International Journal of AdvancedComputer Science and Information Technology (IJACSIT) Vol.2, No. 4, 2013, Page: 56-66, ISSN: 2296-1739.
[15] L. Van Cauwenberge, "Top 10 Machine Learning Algorithms", Data Sci. Cent., 2015.
[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[17] S. Raschka, Python Machine Learning. Unlock deeper insights into machine learn- ing with this essential guide to cutting-edge predictive analytics, Packt Publishing, 2015.
[18] D.G. Altman, Practical Statistics for Medical Research, First Ed., Chapman & Hall, 1990.
[19] "Cleveland heart disease dataset, UCI Repository,1988.http://archive.ics.uci.edu/ml/datasets/Heart+Disease, [Last Accessed 20.01.2020].
[20] https://www.displayr.com/what-is-a-decision-tree/ [Last referred on 18.05.2020].
[21] https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d [Last referred on 18.05.2020].
[22] https://blog.goodaudience.com/machine-learning-using-logistic-regression-in-python-with-code-ab3c7f5f3bed [Last referred on 18.05.2020].
[23] https://medium.com/@sanchitamangale12/voting-classifier-1be10db6d7a5[Last referred on 18.05.2020].

DOI: 10.5373/JARDCS/V12I6/S20201054
*Corresponding Author: Debnath Bhattacharyya, Email: debnathb@kluniversity.in
Article History: Received: Mar 08, 2020, Accepted: June 10, 2020