European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 7, Issue 4, 2020

Random Forest Machine Learning technique to predict Heart disease

Akram Ahmed Mohammed[1], Rajkumar Basa[2], Anirudh Kumar Kuchuru[3], Shiva Prasad Nandigama[4], Maneeshwar Gangolla[5]
[1,2,3,4,5]*Department of Electronics and Communication Engineering, Vidya Jyothi Institute of Technology, Hyderabad, Telangana, 500075, India.*
pearlakram@gmail.com[1], rajkumarbasa51@gmail.com[2], k.anirudhkumarreddy@gmail.com[3], shiva.ng242@gmail.com[4], manishhari1999@gmail.com[5]

Abstract. A Random Forest Machine Learning Algorithm is integrated with the Flask Web framework for predicting of Heart Disease is proposed. The ensemble learning methods are used for predicting heart disease. The proposed methodology involved integration of the Flask Web framework with the Random Forest machine learning technique to estimate the heart disease stages. Artery Blockage indicates the presence of heart disease. The higher the blockage, higher is the stage of heart disease. Stage 1 and Stage 2 indicate the presence of heart disease whereas Stage 3 and Stage 4 are called chronic heart disease and the risk of a heart attack at any day in such patients is very high. The Data required for the prediction contains parameters such as Age, Sex, Blood Pressure, Sugar levels which are collected from the Kaggle website. Experimental results say that predictions by using the proposed approach are consistently better than those obtained using the other methods.

1. Introduction

In our day to day life, individuals are undergoing a routine and busy schedule that ends up in stress and anxiety. Additionally, the proportion of individuals who are weighty and dependent on the role of tobacco goes up drastically. This ends up in diseases like cardiovascular disease, cancer, etc. The difficulty behind these diseases is their prediction. Everybody has completely different values of pulse rate, blood pressure, and sugar levels. However, medically proved the heartbeat rate should be sixty to a hundred beats per minute and the pressure level should be within the range of 120/80 to 140/90.

The survey says a 70% mortality rate is due to heart-related problems. The term heart disease implies various issues that influence the ordinary working of the circulatory framework, which comprises of heart and veins. If the coronary illness is identified at the beginning time and the patient is given proper and sufficient treatment, at that point it tends to be relieved totally, and furthermore, the expense of the treatment can be decreased essentially. So, there is a need to build up an expectation framework to identify the nearness of heart diseases in the patient with higher exactness.

Machine learning algorithms are often used for cardio disease prediction systems. Applying machine learning may be a key approach to utilize massive volumes of accessible Heart-related knowledge. Machine Learning is of nice concern once it involves identification, management, and alternative connected clinical administration aspects. Various machine learning techniques include ensemble classifiers that can be used in improving prediction accuracy. Machine learning techniques help in identifying the data and automatically make the predictions.

The main motivation of the work is to create a web application using the Flask framework and machine learning technique to predict the heart disease stages. The paper sectioned as follows. The Section II deals with the literature that gives the existing methods. The Section III deals with the proposed system. The Section IV deals with experimental results and discussion of the proposed system. The Section V gives the conclusion of the paper.

2. Literature Review

Cardiovascular disease is the main reason for death within the world over the last decade. Nearly one person dies of heart condition concerning each minute within the U. S. alone. To cut back the numbers of deaths from heart diseases there need to be a fast and economical detection technique exploitation data processing. By analyzing the experimental results, it's ended that the J48 tree technique clads to be best classifier for heart condition prediction as a result of it contains a lot of accuracy and the least total time to create [1].

The data mining is referred as the discovery of relationships in giant databases mechanically and in some cases, it's used for predicting the relationships supported by the results discovered. The data mining is compared with reduced variety of attributes. They are named as Naive Bayes, decision Tree and Classification by clustering method.

The Fourteen attributes are reduced to six attributes using genetic search method. Also, the intensity of the sickness supported the results was unpredictable [2]. The foremost effective model to predict patients with cardiomyopathy seems to be a Random Forest classifier. The most challenging in the data mining method or machine learning process is that the inconsistencies of information, presence of missing values, screaming five knowledge and outliers. Therefore, applied math and machine learning methodologies should be applied to manage the information quality [3].

To achieve Smart cardiovascular disease Prediction is constructed via Navies theorem to predict risk factors regarding cardiovascular disease. The present analysis emphasizes on heart condition identification. Information assortment is allotted using various sources that are primary factors accountable for any form of cardiovascular disease and thereby employing a structure the information is made. The speedy advancement of technology has led to an outstanding rise in mobile health technology which is one amongst the online applications [4]. It is impractical for a normal man to often endure expensive tests just like the electrocardiogram and therefore there must be a system in place that is handy and at an equivalent time reliable, in predicting the probabilities of cardiovascular disease. Therefore, we have a tendency to propose to develop an associate application that might predict the vulnerability of a cardiovascular disease given basic symptoms like age, sex, pulse, etc. If the number of individuals using the system will increase, then the notice concerning their current heart standing is identified and therefore the rate of individuals dying because of heart diseases can reduce eventually [6].

Prediction relies on straightforward reasonable medical tests are to be conducted in any native clinic. Moreover, the model is used to determine and supply the additional confidence and accuracy to the Doctor's identification since the model is trained using real-life information of different aged healthy and sick patients. The model is accustomed to assist doctors in analyzing patient as to validate their identification and facilitate decrease human error [7].

Random Forest and Support Vector Machine is applied to make a classifier model that is able to predict sickness with higher performance and accuracy. With current technology in the medical sector, it's potential to cure them with acceptable treatments. However, if it's diagnosed late, then even the high-tech medical instrumentality cannot facilitate [10]. The accurateness of the structure is additionally upgraded by making varied mixtures of information mining techniques and by parameter standardization additionally. This investigation tells the USA regarding dissimilar technologies that are employed in dissimilar papers with a dissimilar count of attributes with completely different accuracies counting on the tools designed for execution [15].

Most of the researchers never targeted on creating a web application and interfacing with the framework. Most developed systems use only algorithms for accuracy and never implemented it in real life with any model which is again insignificant for meaningful information. The data utilized by most researches contain only a few parameters. Hence a suitable model with front end HTML interfacing with the Flask framework through machine learning algorithm should be proposed.

3. Proposed Method

Data Collection is the major step as the quality and quantity of the data that we gather for the proposed system will directly determine how good the results of the predictive model are. We have collected the dataset from the Kaggle (https://www.kaggle.com/datasets). For better and accurate prediction, we consider parameters like Age, Sex, Resting blood pressure (in mm Hg), serum cholesterol (in mg/dl), Fasting blood sugar Rest ECG results, Chest pain during exercise, and Chest pain type. Classification is the task of approximating the mapping function from the input variable to the discrete output variable. It is the classification where the outcome is either true or false (1 or 0). Random forest is an estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. In the regression model, the prediction is based on the independent variable. Random forest in regression operates on constructing a multitude of decision trees at training time and outputting the class that is the mode of mean prediction of the individual trees. After reading the data the data visualization method is performed. In this method, the large data sets are transformed into a statistical and graphical representation.

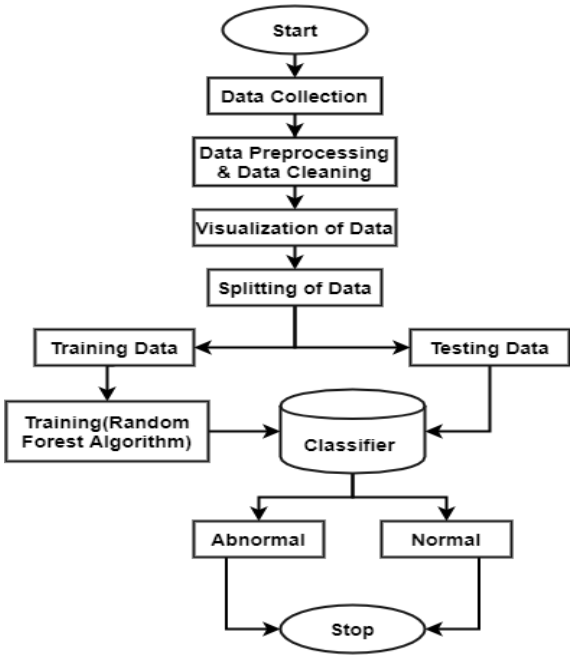

Fig. 1. Flow Chart

Data required for the prediction is collected using open resources. Data collection is an important step as the quality and quantity of the data that we gather for the proposed system will directly determine how good output that predictive model can be. The general approach is that we can collect data from Open sources like Kaggle (https://www.kaggle.com/datasets) For better and accurate prediction, we consider parameters like Cholesterol, Age, Blood Pressure, Chest Pain, Sex, Fasting Blood Sugar, Rest ECG, and Heart Rate during ECG.

Data processing is defined as the collection and manipulating of data to produce the desired meaning and understandable data for prediction. In this stage regression and classification, techniques are used for the process. The main steps that included in data processing are as follows:
Import the libraries and datasets.
Data cleaning.

Data Transformation.

Data visualization is the method of transforming large data sets into a statistical and graphical representation. It is an important task of data science and these are the useful techniques that make the data less confusing and more accessible. Data Visualization takes a huge complex amount of data and then represents them in the form of charts or graphs for better understanding. Analysis using a Heat map, bar graphs, and pair plots are mentioned below

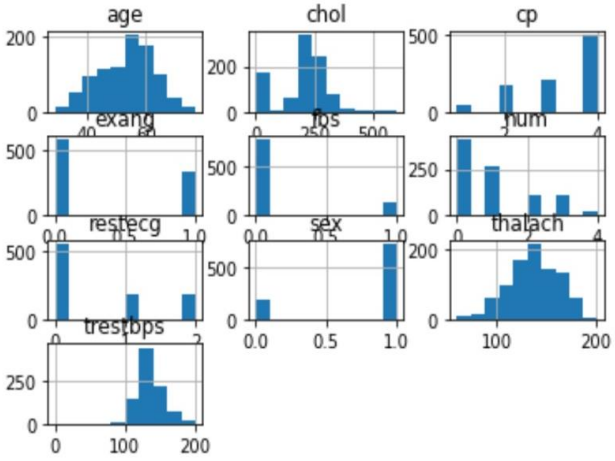

Fig. 2. Visualization of parameters related to heart disease

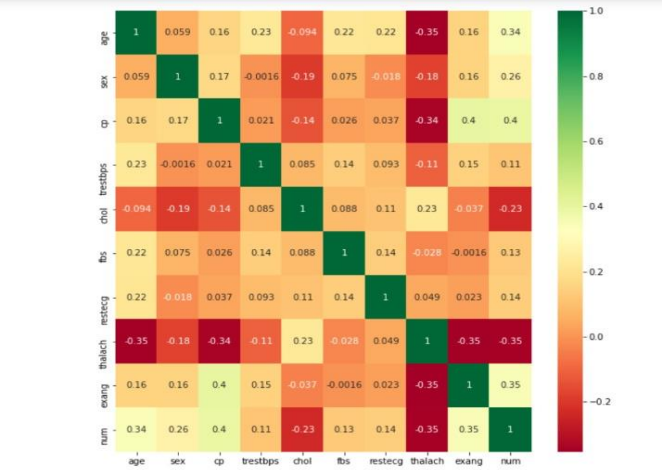

Fig. 3. Heat map of correlated features

The splitting of data is classified into training and testing of data. Training is applied to the 75 parts of the data set. Testing is applied to the 25 parts of the dataset. Testing the data is used to evaluate the performance of the model using a few algorithms. Based on the training data and testing data the best model is selected. The training data is different from testing data; the obtained data is applied to the algorithm. The flowchart of the proposed methodology is shown in figure 1.

The work aims to predict the diagnosis of heart disease and its stages. Different classifiers and regression models have been used in the paper but very few models listed below gives less error. Random Forest is a supervised learning algorithm. It is an ensemble learning method for classification, regression, and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the class or mean prediction of the individual trees. This algorithm creates decision trees on data samples. There is a direct relationship between the number of trees in the RF and the results it can get. After creating the decision trees, it gets the prediction from each of them (tree), and then finally it selects the best solution employing voting. The larger the number of trees, the more would be the accuracy of the result. The larger the trees the better will be accuracy.

The main advantage of using Random forest is that it has high accuracy and less variance than a single decision tree. The Data Pre-processing is done and then based on the parameters in the dataset; the number of trees formed as shown in figure 4. The accuracy of the result directly depends on the number of trees, so if the number of trees is more the accuracy would be high. In the proposed model, we split the data into 75% training data and 25% testing data. Accordingly, the data is trained and tested in all possible combinations and finally, it gives the best model. Now, the model is trained with the help of the Random Forest Classifier. From each tree, we get a predicted result and have entropy which is calculated individually. Voting (which

includes combining the predictions) is performed for all the predicted results and finally, the most voted prediction result is selected.

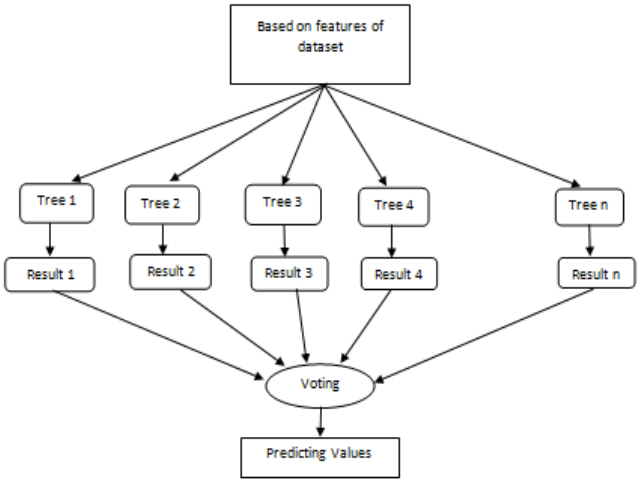

Fig.4 Random Forest Illustration

4. Results and Discussion

Our work is meant to predict heart disease diagnosis. The required parameters are taken as input and are processed by the random forest algorithm techniques integrated with the Flask framework. A web page has been created with the desired parameters as input with HTML front end technology. After the completion of the processing of input data of an individual, the results are displayed and categorized into four stages based on the input data values of an individual and diagnosed in different stages. These stages are classified based on the parameters involved in heart disease prediction such as blood pressure levels, cholesterol levels, age, and sex, etc. Blood pressure is further classified as hypertension or high blood pressure and hypotension or low blood pressure. Similarly, cholesterol levels are classified into high-density lipoproteins and low-density lipoproteins. These values decide the stages of an individual and predict the result.

Heart disease can be classified into five stages (stage 0 to 4) based on severity of artery blockage. Artery blockage >50% indicates presence of heart disease. Higher the blockage, higher is the stage of heart disease. Stage 0 indicates no heart disease. Stage 1 and Stage 2 indicates the presence of heart disease whereas Stage 3 and Stage 4 are called as chronic heart disease and risk of heart attack at any day in such patients is very high. The images of the results are shown below

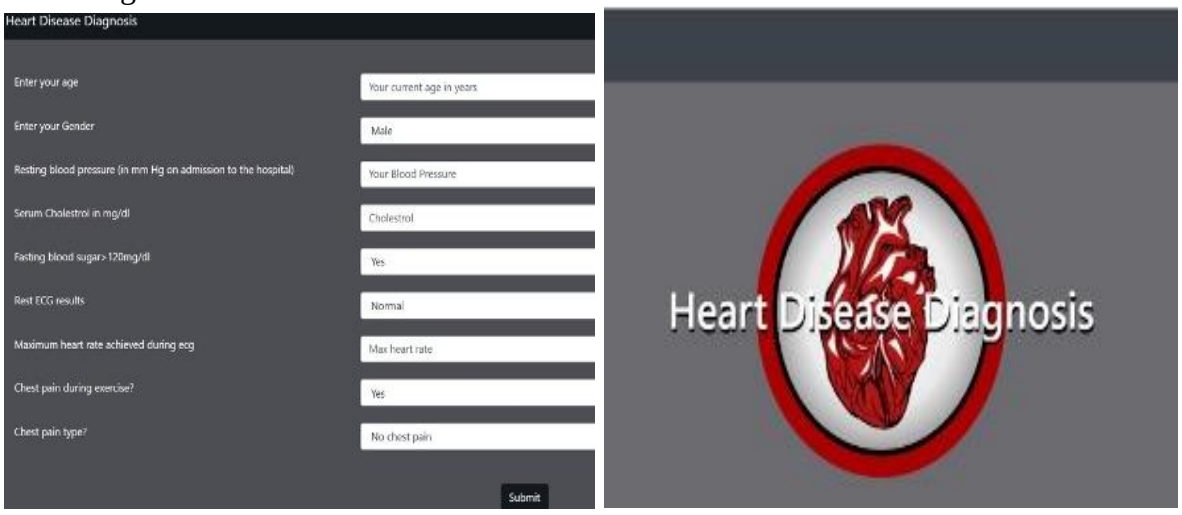

Fig.5 Introduction of Home and About page

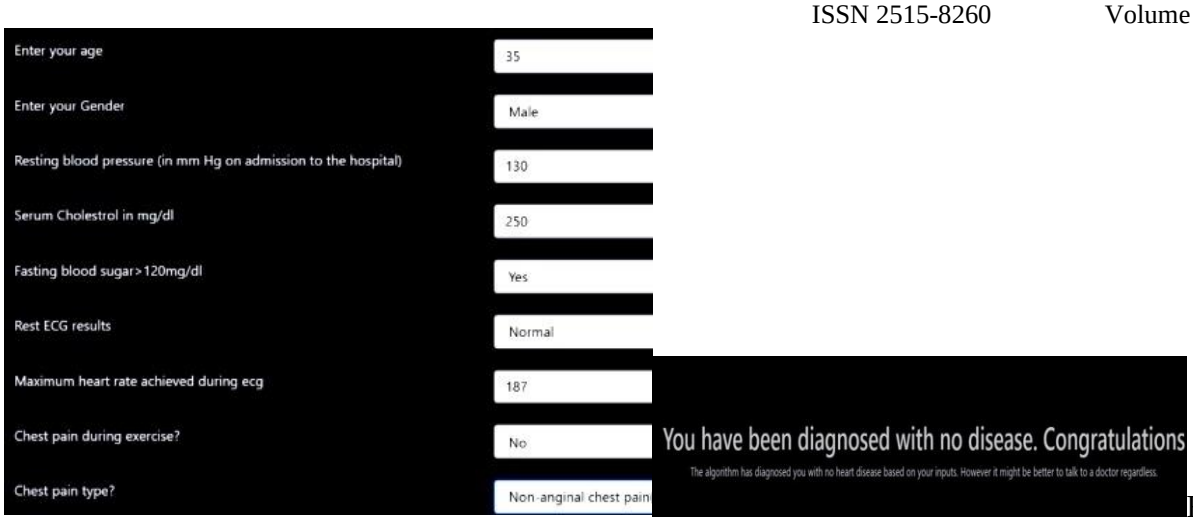

Fig.6 Stage 0 Input and Output details

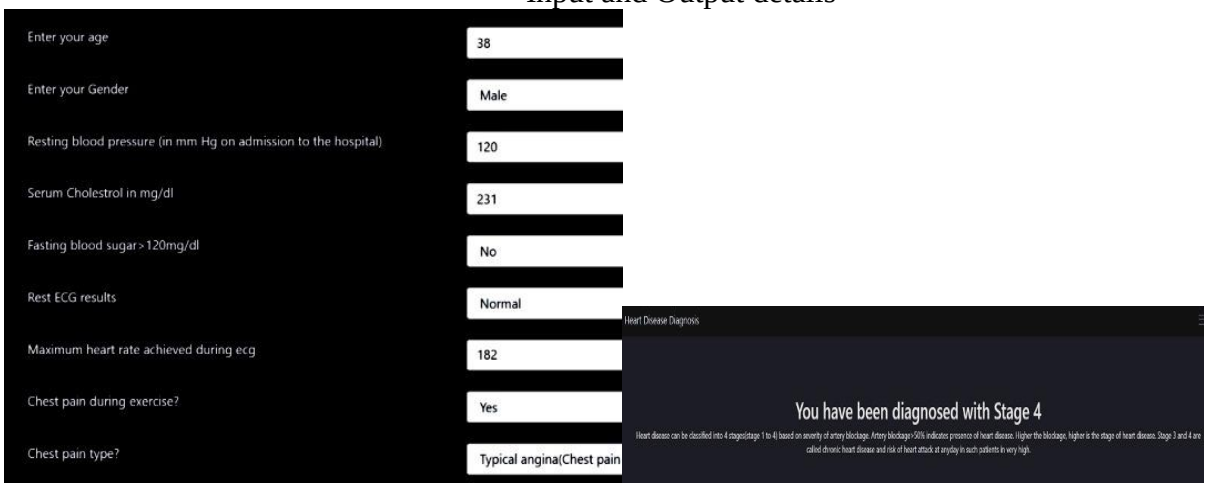

Fig.7 Stage 4 Input and Output details

If the person takes precautions at Stage 1 and Stage 2 then the person has more chances to recover. As Stage 3 and Stage 4 are critical and probability of getting heart disease recover rate is less so the person should consult to doctor and should be more careful.

5. Conclusion and Future scope

This paper proposes a random forest algorithm model integrated with Flask framework and HTML front end technology for the prediction of heart disease diagnosis. Thus, the proposed model gives the best predicted values with high accuracy. If we are unable to identify this disease at the initial stage then the chances of curing this disease is very critical. Our further goal is to extend this research by a real-time system using Deep Learning approach, where users can upload their test results as image.

6. References

[1] Jaymin Patel, Prof.Tejal Upadhyay, Dr. Samir Patel (2016). Heart Disease Prediction Using Machine Learning and Data Mining Technique. In International Journal of Computer Science & Communication (IJCSC), Volume-7, Number 1 Sept 2015- March 2016 Page No.129 – 137. Shi, D.-P.,Wu, C. “The influence of infrared temperature measurementbased on reflection temperature compensation and incidenttemperature compensation” Electron. Measur. Technol. 08, 2321–2326 (2015)

[2] Shamsher Bahadur Patel, Pramod Kumar Yadav, Dr. D. P.Shukla (2013). Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques. In IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS) E - ISSN: 2319-2380, p-ISSN: 2319-2372. Volume 4, Page No. 61-64.

[3] SanchayitaDhar, Krishna Roy, Tanusree Dey, Pritha Datta, Ankur Biswas(2018).A Hybrid Machine

Learning Approach for Prediction of Heart Diseases. In 2018 4th International Conference on Computing Communication and Automation (ICCCA).

[4] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin(2019). Design And Implementing Heart Disease Prediction Using Naives Bayesian. In Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8.

[5] Shanmugasundaram G, Malar Selvam V, R. Saravanan(2018). An Investigation of Heart Disease Prediction Techniques. In Conference at SMVEC.

[6] Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Kailas Devadkar(2018). Prediction of Heart Disease Using Machine Learning. In Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018).

[7] Rahma Atallah, Amjed Al-Mousa(2019). Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method.In Conferenece at Princess Sumaya University for Technology.

[8] Nidhi Bhatla , Kiran Jyoti(2012). An Analysis of Heart Disease Prediction using Different Data Mining Techniques. In International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 8, October - 2012 ISSN: 2278-0181.

[9] Ayantan Dandapath, M Karthik Raja(2018).Heart disease prediction using machine learning techniques a survey. In International Journal of Engineering & Technology IJET.

[10] Akshay Jayraj Suvarna, Arvind Kumar M, Ajay Billav, Muthamma K M, Gadug Sudhamsu(2019). Predicting The Presence of Heart Disease Using Machine Learning.In International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 8, Issue. 5, May 2019, pg.119 – 125.

[11] Abhijeet Jagtap, Priya Malewadkar, Omkar Baswat3,Harshali Rambade4(2019). Heart Disease Prediction using Machine Learning.In International Journal of Research in Engineering, Science and Management Volume-2, Issue-2, February-2019).

[12] Gnaneswar B, Ebenezar Jebarani M.R.(2017). A Review on Prediction and Diagnosis of Heart Failure. In 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS).

[13] Md. Jamil-Ur Rahman1, Rafi Ibn Sultan2, Firoz Mahmud3, Ashadullah Shawon4 and Afsana Khan5(2018). Ensemble of Multiple Models For Robust Intelligent Heart Disease Prediction System. In 4th International Conference on Electrical Engineering and Information & Communication Technology.Sudharsan, RR, J. Deny, E. Muthukumaran, and R. Varatharajan. “FPGA based peripheral myopathy monitoring using MFCV at dynamic contractions.” Journal of Ambient Intelligence and Humanized Computing, 1-9 (2020).

[14] Purushottam , Prof. (Dr.) Kanak Saxena, Richa Sharma(2016). Efficient Heart Disease Prediction System. In Procedia Computer Science 85 (2016) 962 – 969.

[15] Avinash Golande, Pavan Kumar T(2019). Heart Disease Prediction Using Effective Machine Learning Techniques.In International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019.