A Study on Risk Prediction of Cardiovascular Disease Using Machine Learning Algorithms

[1]**MALKARI BHARGAV**, [2]**J.RAGHUNATH**
[1]M.Tech Student, [2]Assistant Professor,
Department of Computer Science and Engineering,
Gates Institute Of Technology, Affiliated to JNTUA, Gooty, India.

Abstract : In this modern days, there are many changes occur in our daily life. It will mainly impacting on health system. As a result of this various changes, health diseases are rapidly increasing in our day to day life. Here is some of diseases are more affected in our life. Such as cardiovascular Diseases, Stress Depression, Cancer and many more diseases are present in our today's life. Mainly, cardiovascular disease is more commonly affected in our life. It will affect in any age group persons. The main cause of this cardiovascular disease is changes in the Blood Pressure, Cholesterol, increasing Heartbeat etc.. It may lead to risk for life and death also. Coronary Heart disease is caused by fatty plaque deposits on narrowed arteries walls supply to the heart and it will reduce the flowing of blood in heart. The main aim of this project is to predict the heart disease with machine learning algorithms and diagnose in early stages. In this research, we are implementing different machine learning algorithms with UCI dataset to find the best accuracy in different algorithms. Then i got best accuracy in Artificial Neural Network. So ANN classification algorithm is used to know the possibilities of getting heart disease and diagnose in initial stage.

***Index Terms* - Support Vector Machine, Decision Tree, Logistic Regression, Random Forest, KNN, Artificial Neural Network, Machine Learning, Keras, Tensor Flow, Cardiovascular Disease Prediction.**

I. INTRODUCTION

Nowadays, there are so many diseases affected in our life. Mainly Cardiovascular heart disease is most commonly affected in our human life. And these are the most common types of cardiovascular diseases is cardiac arrest, hypertension, Heart strokes, coronary Heart disease and many more. Coronary heart disease can be identified with different medical test and diagnosed it. But medical tests are taking more time to detect and quite difficult. Sometimes it leads to mortality. According to World Health Organization, cardiovascular disease is main cause of deaths in worldwide. Around 17.9 million people are mortality due to cardiovascular disease in the year 2016. Out of these 85% deaths occurs due to heart stroke and heart failure [1] and in India also around 1.7 million peoples are died due to Heart disease [2]. The rate of mortality are rapidly increased due to changes of human behavior habits like smoking, stress depression and drinking alcohol [4] in day to day life. If we predict the heart disease in early stage, lot of patients can take proper treatment and prevention from these diseases without taking any risk to the life. And it will automatically decrease the rate of mortality and increase the recovery rate. So mainly prediction of heart disease are quite difficult in early stages and it plays crucial role to predict. Heart is the most important organ in the human body.

In this days, it is important to develop the medical diagnosed system to identify the possibilities of getting heart disease in a short time. So that, we can prevention the heart disease on early stage, and we can decrease the risk from heart disease and death. When we talking about diagnose system, the machine learning(ML) algorithm are most popular technique to predict the risk of heart disease with maximum accuracy in less time[3]. In machine learning algorithms, artificial neural network is the one of the best machine learning technique which is used to predict the possibilities of heart disease with best accuracy and efficient results in a short period. In this proposed system, we are using different algorithms has been implemented to predict the risk of heart disease. In this research, we are used UCI dataset to implement the prediction of heart disease by using different machine learning techniques to get best accuracy. In the further studies we are discussed about II. Literature survey III. Proposed system IV. Data source V. Methodology VI. Result VII. Conclusion and References.

II. LITERATURE SURVEY

In this contemporary days, Machine learning algorithm are most important and popular techniques to use future analysis prediction like stock exchange, data analysis and so on. In medical field also machine learning playskey role to predict the risk of any possibility of diseases are affecting in patients. In this section we discuss about various researchers are used machine learning techniques to analysis the medical database as follows.

S,Mondal [5] is proposed to identify the heart disease by using mobile application with data mining technique. He used 917 patients record with 10 attributes to predict the risk using decision tree and he got 86 percent accuracy. In another research Parichay Kumar Mandal [6] had been proposed to identify the ischemic heart disease by using back propagation algorithm for ANN. He is collected the some patient data with 14 features and he achieved 84% accuracy. Thanigaivel [7] is used the data mining technique to prediction the heart disease system. In his proposed system he used decision tree and got accuracy 68 percent. Monika Gandhi [8] made use to forecast the heart attack with neural network, Naive Bayes and decision tree. In this research she collected more amount of data is used to identify the disease. M.Marimuthu[9] used some machine learning algorithms like support vector machines, k nearest neighbor, decision tree and Naive Bayes used to analysis the heart disease and he achieved the accuracy like 65%, 83%, 75%, 80% among all these he got best accuracy in Knn algorithm. Shashikant [10] used two classifiers are feed forward algorithm and support vector machine technique to diagnosis the disease. He obtained the best accuracy as 86 percent and 85 percent as sensitivity in SVM classifier. R.das [11] the back propagation algorithm of neural network is used to identify the heart disease and he obtains 89% accuracy. In his proposed research, the achieved results are compares with existing system with same neural network research and achieve improved results in accuracy.

In the existing system, some of disadvantages are obtained in machine learning techniques. Many researchers are used decision tree technique. There are some drawbacks in this technique are sometimes it takes so much time to train and calculation the model compared with other machine learning techniques. Some researchers are used SVM technique but here also some drawback are svm doesn't work on large dataset very well and sometimes it overlaps the classes. In knn technique it doesn't learn the any data before training the model and continues to training the model. Some researchers are uses large data with less attributes to predict the disease.

III . PROPOSED SYSTEM

In our research, we are using different algorithms to get best accuracy from other existing system results. So we are trying to predict the cardiovascular heart disease with best accuracy by using different machine learning algorithms are implemented [fig.3].They are:

1. Random Forest
2. K-nearest Neighbor
3. Support Vector Machine
4. Logistic Regression
5. Decision Tree
6. Artificial Neural Network

IV. DATASET SOURCE

In this study, we are utilize about cardiovascular dataset from Cleveland Hospital databases. And we are collected the dataset is available in UCI Machine learning repository [13]. The dataset carries 303 patients samples with 14 parameters like age, cholesterol, blood pressure and so on. The details of parameters [table 1] and sample data [fig.2]. In fig.1 we are shown how many are affected with heart disease in dataset.

Table 1. Dataset Parameters

S.NO	PARAMETERS	DATA TYPE
1.	Age	Integer
2.	Sex	Integer
3.	Chest Pain	Integer
4.	Blood Pressure	Integer
5.	Cholesterol	Integer
6.	Fbs(sugar)	Integer
7.	ECG(Electrocardiographic)	Integer
8.	Thalach (Maximum heart Rate)	Integer
9.	ST Depression(Oldpeak)	Float
10.	Slope	Integer
11.	Exercise angina(exang)	Integer
12.	Ca(Number of vessels)	Integer
13.	Thalassemia	Integer
14.	Target	Integer

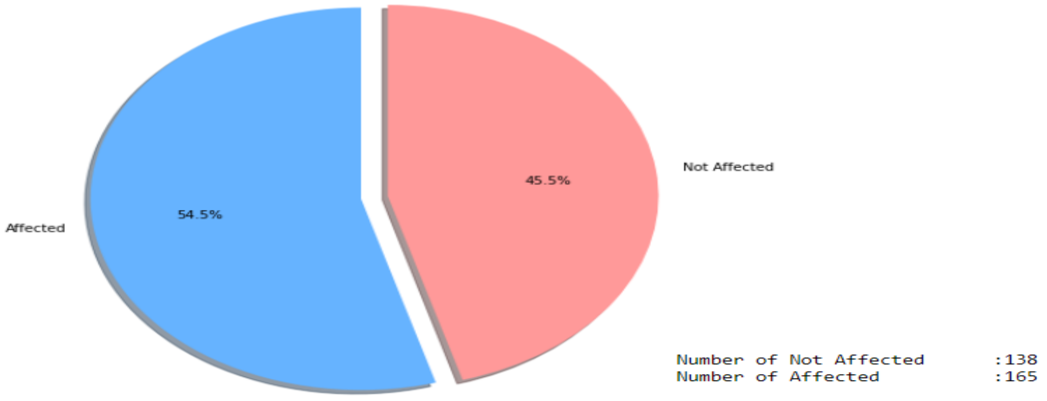
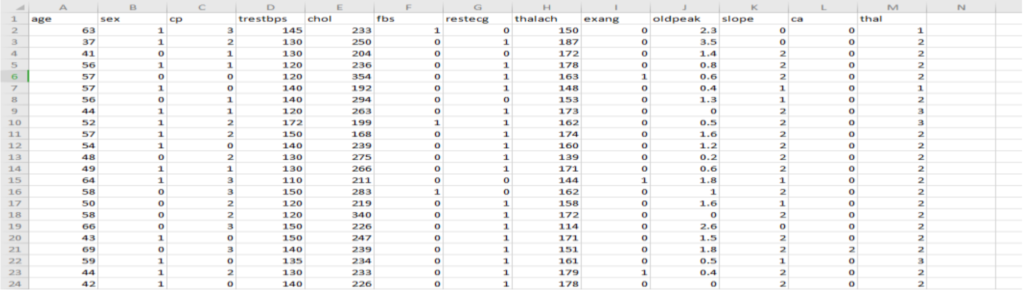


Fig.1. Pie Chart



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	
2	63	1	3	145	233	1	0	150	0	2.3	0	0	1	
3	37	1	2	130	250	0	1	187	0	3.5	0	0	2	
4	41	0	1	130	204	0	0	172	0	1.4	2	0	2	
5	56	1	1	120	236	0	1	178	0	0.8	2	0	2	
6	57	0	0	120	354	0	1	163	1	0.6	2	0	2	
7	57	1	0	140	192	0	1	148	0	0.4	1	0	1	
8	56	0	1	140	294	0	0	153	0	1.3	1	0	2	
9	44	1	1	120	263	0	1	173	0	0	2	0	3	
10	52	1	2	172	199	1	1	162	0	0.5	2	0	3	
11	57	1	2	150	168	0	1	174	0	1.6	2	0	2	
12	54	1	0	140	239	0	1	160	0	1.2	2	0	2	
13	48	0	2	130	275	0	1	139	0	0.2	2	0	2	
14	49	1	1	130	266	0	1	171	0	0.6	2	0	2	
15	64	1	3	110	211	0	0	144	1	1.8	1	0	2	
16	58	0	3	150	283	1	0	162	0	1	2	0	2	
17	50	0	2	120	219	0	1	158	0	1.6	1	0	2	
18	58	0	2	120	340	0	1	172	0	0	2	0	2	
19	66	0	3	150	226	0	1	114	0	2.6	0	0	2	
20	43	1	0	150	247	0	1	171	0	1.5	2	0	2	
21	69	0	3	140	239	0	1	151	0	1.8	2	2	2	
22	59	1	0	135	234	0	1	161	0	0.5	1	0	3	
23	44	1	2	130	233	0	1	179	1	0.4	2	0	2	
24	42	1	0	140	226	0	1	178	0	0	2	0	2	

Fig. 2. Sample data

V. METHODOLOGY

The main aim of this project is to diagnose the cardiovascular disease with the help of machine learning algorithms and it will be the fastest way to predict and cure the disease at right time. We are showing different machine techniques to diagnose the heart disease with some selected parameters are tested. In this methodology we are structured in different stages to predict [Fig.3].

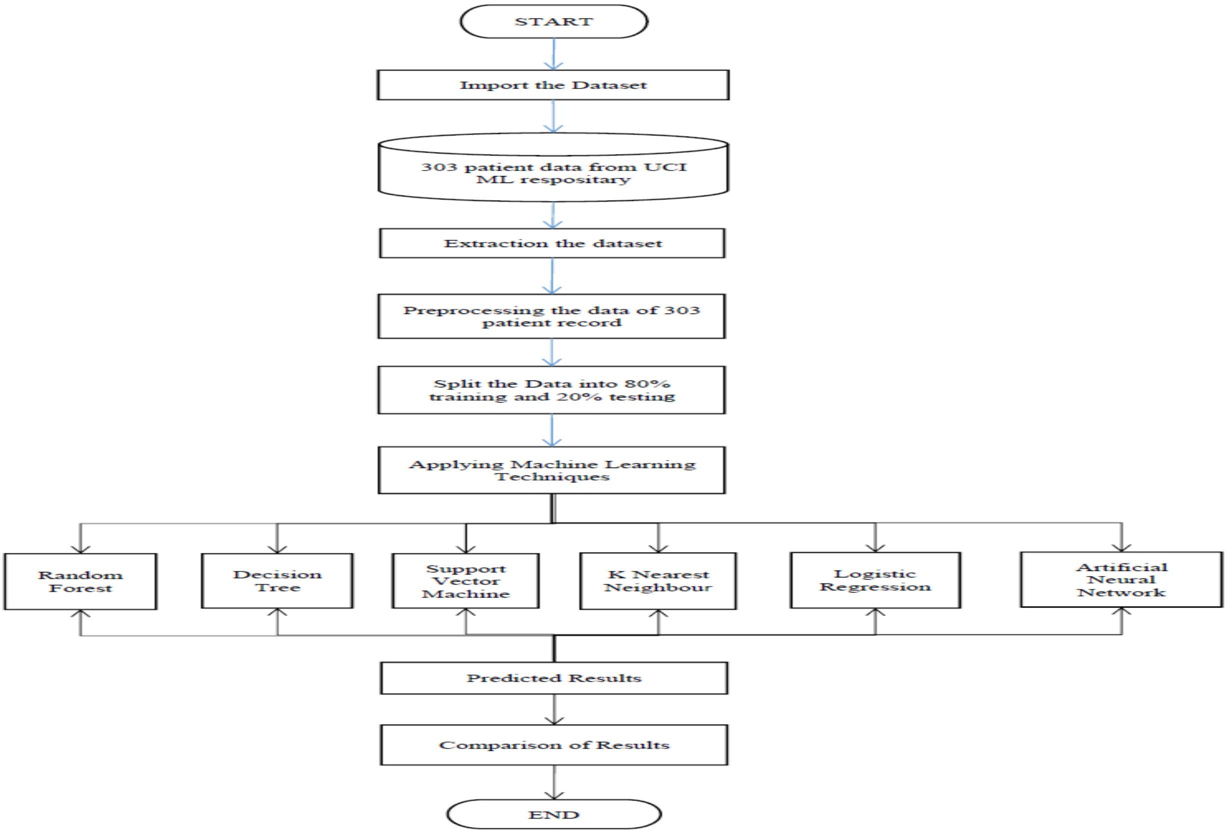

START
Import the Dataset
303 patient data from UCI ML respositary
Extraction the dataset
Preprocessing the data of 303 patient record
Split the Data into 80% training and 20% testing
Applying Machine Learning Techniques
Random Forest
Decision Tree
Support Vector Machine
K Nearest Neighbour
Logistic Regression
Artificial Neural Network
Predicted Results
Comparison of Results
END

Fig.3. Work flow of proposed system

In this proposed system, we are used three main various tools and different machine learning algorithms.

1) Tools:

- **Tensor Flow:**It is the most popular machine learning library in the globe. Google artificial intelligence organisation and brain team are developed tensor flow library. It is used to speed up the machine learning and AI system. It can use in different language like python, java[12]
- **Keras:** It is an application programming interface designed for user. It is frontend library for Tensor flow. Keras and tensor flow is jointly works as one tool.[12]
- **Jupyter Notebook:** It is open source application are available in anaconda. It is easy to create any python code and data visualization. In anaconda, there are nearly 1400 packages are already available in environment in the repository.

2) Algorithms:

2.1. Random Forest:

Random forest is a one of the best algorithm in machine learning and it is member of Supervised Learning. Random forest runs same as Essemble learning. It is nothing different that more decision trees are collected in one forest like that only it will solves the problem with best accuracy of these model. In this algorithm, a tree can splits into nodes and it predicts the each node are available in tree.

The main advantage of random forest algorithms are it is very easy to solve the algorithm with good accuracy and it can able to handle the large number of datasets. It is proficient in both the classification and regression.

2.2. Decision Tree:

Decision tree is another type of machine learning algorithm and it is also comes under supervised learning. It uses the structured type of tree to solve the predictive model. In which each tree having parent and child nodes to represents a structured tree.

The main advantage of Decision Tree algorithms are it takes less time and less attempt for pre-processing the data in a model comparing with other algorithms in a machine learning. It is very attractive and easy to explain the model to clients.

2.3. Support Vector Machine:

Support vector machine is the one of the most popular technique in supervised machine learning algorithm and it will learns the data for classification and regression. In this algorithm, there is training phase to train the given data in beginning stage. After completing training phase then it will builds the model and it finds which classifiers are represented with the help of human. This step also called as feature extraction. In SVM model, we are presentation the samples as a points in space. Different samples are separated by points and mapped also. In between the points there is a line is called hyper plane. A hyper plane line is represents as divided line and it separates the points.

The main advantage of Support Vector Machine algorithm are it will work quietly on unstructured and semi structured data. It represents the clear line in between points and it has less risk to over fitting the model.

2.4. K-Nearest Neighbor:

K-Nearest Neighbor is one of simplest and easiest technique in machine learning. In this algorithm, the available data can process to predict the output. In case when we input new data in the classifier it can run the new data as same as available old data stored in classifier. It can predict the output along with new data added in classifier. When we increase the new data. Then the k value also increase the increase.

The main advantage of K-Nearest Neighbor algorithms are it will more successful, when the large number of data are available. It is simple and easy way to predict the output.

2.5. Logistic Regression:

It is another technique in machine learning algorithm. It is go method for binary classification problems. It is used the coefficient model from database and it is used to prediction the probability of result. In this Logistic Regression, it is simply uses the binary value as 1(yes) or 0(no). Mainly this technique is uses for prediction of diseases and detection. It is also known as logistic function. The main advantage of Logistic Regression algorithms it is easy to apply and train the model.

2.6. Artificial Neural Network:

Artificial neural network (ANN) is one of the algorithm in deep learning. Deep learning is a sub group of machine learning. It is a supervised learning. ANN model is behaves like a human brain. It can simulate the network of neurons like a nervous system in brain. It can learn the data and analysis the network then it automatically makes own decision as same as human thinking without in any interfere of humans. ANN have capability to solve the impossible problems. In human brain, there are so many millions of nerves attached with any body parts present in human body. In ANN also there are millions of neurons are interconnected by layers. The architecture of ANN is to connect the neurons in 3 layers. Input layer collects the data as neuron from outside and then it sends to another layer. Hidden layer in between the two layers and it works as prepossessing the neurons and it sends to final layer. Output layer is to produce the result (Fig. 4). In this research we are used multi-layer neural network. In this multi-layer network there are one or two hidden layers are available. So that it have capability to solve impossible problems in this network.

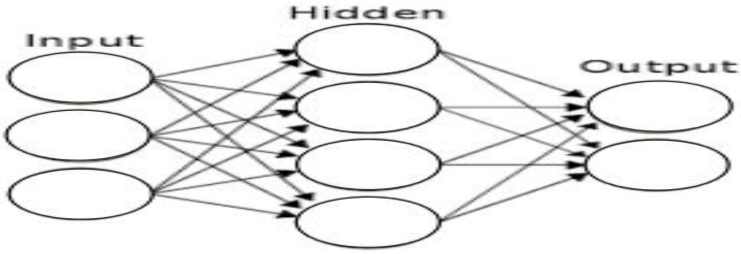


Fig. 4. Architecture of ANN Layers

In machine learning, the classification report is used to calculate the standard of predicted values. With this report we can get four types of results as precision, f1 score, recall, support values (Table.2). With this results we can calculate the f-measures, specificity, and sensitivity. So that we may estimate the predicted result is true or false.

VI. RESULT :

In this study, we are showing the Classification report of different Machine Learning algorithms using UCI dataset.

A) Classification Report:

1) Precision = true positive / (true positive + false positive)

2) F-measure = 2*true positive / (2*true positive + false positive + false negative)

3) Sensitivity = true positive / (true positive + false negative)

4) Specificity = true negative / (true negative + false positive)

Table 2. Classification Report

Classifications / Algorithms	Precision	F-Measure	Sensitivity	Specificity
Artificial Neural Network	94.29	97.29	100.0	85.71
Logistic Regression	87.87	89.23	90.62	85.20
Random Forest	84.37	84.37	84.37	82.75
Decision Tree	89.28	83.33	78.12	89.65
Support Vector Machine	66.66	75.67	87.5	51.74
K-Nearest Neighbor	68.57	71.64	75.0	62.06

B) ACCURACY REPORT:

In below, we are showing accuracy report of different ML algorithms using UCI dataset.

1) Accuracy = (true negative + true positive) / (true negative + true positive + false negative + false positive)

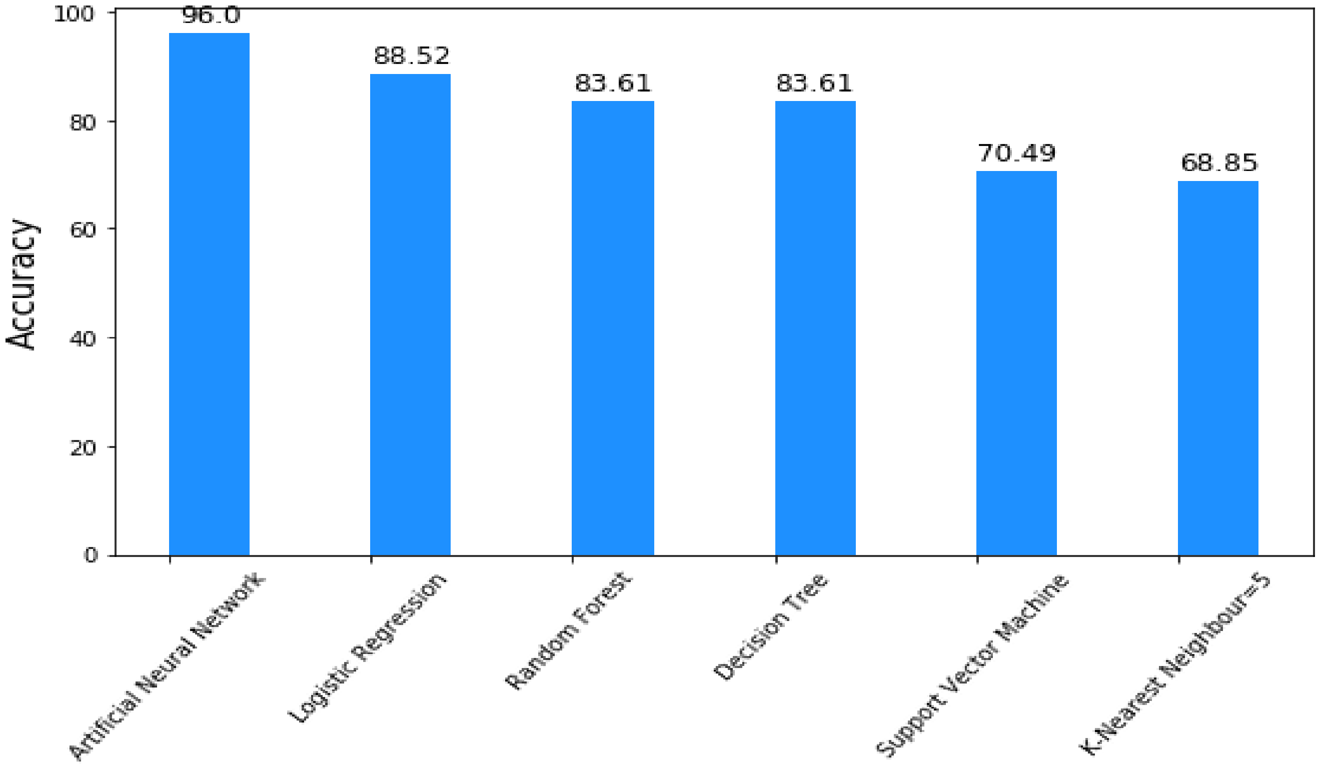


Fig.5. Accuracy Report

Finally, In this Research Paper, We got highest accuracy in Artificial Neural Network with 96.00% accuracy using dataset [fig. 5]. In UCI dataset classification report, the Precision, F-Measure, Sensitivity, Specificity are 94.73%, 97.23%, 100% and 85.71% respectively[Table.2] with Artificial Neural Network algorithm. And second highest accuracy got in Logistic Regression with 88.52% accuracy with dataset. So in this research ANN is more accurate to prediction the Heart disease in early stages.

VII. CONCULSION :

Mainly In this paper, we are implemented six Machine learning techniques are SVM, KNN, Random Forest, Logistic Regression, Decision Tree, ANN algorithms to find the risk of heart disease prediction with best accuracy. In this research, we are used nearly 303 patients record from UCI dataset. We used the dataset and trained with all the ML algorithms separately. After that, we are achieved best accuracy and result with artificial neural network algorithm in UCI dataset are 96.00% and followed by logistic regression are 88.52% respectively. In this proposed system, we got artificial neural network is better method than the remaining algorithms. Finally, the artificial neural network is best technique to find the possibilities of cardiovascular disease by using different parameters in early stages. In Future study, researcher can use some advance deep learning algorithms like Conventional neural network, Long short-term memory, Generative Adversarial Network, RNN etc.,. To identify the heart disease with maximum accuracy using more number of parameters.

VIII. REFERENCE :

1. “Cardiovascular diseases”, World Health Organization, 2020. [Online]. Available:https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases (cvds) [accessed: 15-june-2020].
2. “Heart disease in India “Global Burden of disease, 2017. [Online]. Available: http://www.healthdata.org/india [accessed: 15-june-2020]
3. S. Singh, S.Bharti, “Analytical study of heart disease prediction comparing with different algorithms”, International Conference on Computing, Communication & Automation, 2015, IEEE
4. T. Karaylan and. Kl, “Prediction of heart disease using neural network”, International Conference on Computer Science and Engineering (UBMK), 2017, IEEE.
5. M. Raihan, S. Mondal, A. More, P. Boni, “Smartphone Based Heart Attack Risk Prediction System with Statistical Analysis and Data Mining Approaches”, Advances in Science,Technology and Engineering Systems Jaurnal, 2017.
6. Parichay Kumar Mandal, Muhammad Muinul Islam, Tanvir Hossain, Promila Ghosh, Shekel Ahmed Shaj, Abdullah Anik, Mubtasim Rafid Chowdhury, M. Raihan, Saikat Mondal and Arun More “Risk Prediction of Ischemic Heart Disease Using Artificial Neural Network” Internatianal Conference on ECCE, 2019, IEEE.
7. R. Thanigaivel, and K. Ramesh Kumar. "Boosted Apriori: an Effective Data Mining Association Rules for Heart Disease Prediction System." Middle-Eost Journal of Scientific Research, 2016.
8. Monika Gandhi, Shailendra Narayanan Singh, “Predictions in heart disease using techniques of data mining”, International Conference on futuristic trends on computational analysis and knowladge management, 2015, IEEE.
9. M. Marimuthu, S.Deivarani, Gayathri.R,”Analysis of Heart Disease prediction using various Machine LearningTechniques”, International Conference on Artificial Intelligence, Smart Grid and Smart City Applications, 2019, IEEE.
10. Shashikant Ghumbre, Chetan Patil, Ashok Ghatol, “Heart Disease Diagnosis using Suport Vector Machine”, Internatianal C0nference on Information Technology, 2011, ICCIT.
11. Resul Das, Ibrahim Turkoglu, and Abdulkadir Sengur, ``Effective diagnosis of heart disease through neural networks ensembles", Expert System with Applications: An Internati0nal Journal, 2009.
12. “Tensor Flow”[Online].Available: https://www.tensorflow.org/ [accessed: 15- June- 2020].
13. “UCI dataset” [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+Disease/ [accesed: 15-febuary-2020].