



# Heart Disease Prediction using Machine Learning

Sadiq Jaffer<sup>1</sup>, Sameer Pasha<sup>1</sup>, Shailaja L K<sup>2</sup>

Student Dr. Ambedhkar Institute of Technology Bengaluru, India<sup>1</sup>

Associate Professor Dr. Ambedhkar Institute of Technology, Bengaluru, India<sup>2</sup>

**Abstract:** In recent times, Heart Disease prediction is one of the most complicated tasks in medical field. In the modern era, approximately one person dies per minute due to heart disease. Data science plays a crucial role in processing huge amount of data in the field of healthcare. As heart disease prediction is a complex task, there is a need to automate the prediction process to avoid risks associated with it and alert the patient well in advance. This paper makes use of heart disease dataset available in UCI machine learning repository. The proposed work predicts the chances of Heart Disease and classifies patient's risk level by implementing different data mining techniques such as Decision Tree, Knn, Logistic Regression and Random Forest. Thus, this paper presents a comparative study by analyzing the performance of different machine learning algorithms. The trial results verify that Random Forest algorithm has achieved the highest accuracy of 90.16% compared to other ML algorithms implemented.

**Keywords:** Decision Tree, KNN, Logistic Regression, RandomForest, Heart Disease Prediction

## I. INTRODUCTION

The work proposed in this paper focus mainly on various data mining practices that are employed in heart disease prediction. Human heart is the principal part of the human body. Basically, it regulates blood flow throughout our body. Any irregularity to heart can cause distress in other parts of body. Any sort of disturbance to normal functioning of the heart can be classified as a Heart disease. In today's contemporary world, heart disease is one of the primary reasons for occurrence of most deaths. Heart disease may occur due to unhealthy lifestyle, smoking, alcohol and high intake of fat which may cause hypertension [2]. According to the World Health Organization more than 10 million die due to Heart diseases every single year around the world. A healthy lifestyle and earliest detection are only ways to prevent the heart related diseases. The main challenge in today's healthcare is provision of best quality services and effective accurate diagnosis [1]. Even if heart diseases are found as the prime source of death in the world in recent years, they are also the ones that can be controlled and managed effectively. The whole accuracy in management of a disease lies on the proper time of detection of that disease. The proposed work makes an attempt to detect these heart diseases at early stage to avoid disastrous consequences.

Records of large set of medical data created by medical experts are available for analyzing and extracting valuable knowledge from it. Data mining techniques are the means of extracting valuable and hidden information from the large amount of data available. Mostly the medical database consists of discrete information. Hence, decision making using discrete data becomes complex and tough task. Machine Learning (ML) which is subfield of data mining handles large scale well-formatted dataset efficiently. In the medical field, machine learning can be used for diagnosis, detection and prediction of various diseases. The main goal of this paper is to provide a tool for doctors to detect heart disease as early stage [5]. This in turn will help to provide effective treatment to patients and avoid severe consequences. ML plays a very important role to detect the hidden discrete patterns and thereby analyze the given data. After analysis of data ML techniques help in heart disease prediction and early diagnosis. This paper presents performance analysis of various ML techniques such as Naive Bayes, Decision Tree, Logistic Regression and Random Forest for predicting heart disease at an early stage [3].

## II. RELATED WORK

Lot of work has been carried out to predict heart disease using UCI Machine Learning dataset. Different levels of accuracy have been attained using various data mining techniques which are explained as follows.

Some papers were published around 2 to 3 years back have a less accuracy for the prediction of heart disease as compare to today's need. 'Efficient heart disease prediction system using decision tree by sharma et al. it was published in 2015. they have used a decision tree classifier with getting accuracy 75.

T. Nagamani, et al. have proposed a system [2] which deployed data mining techniques along with the MapReduce algorithm. The accuracy obtained according to this paper for the 45 instances of testing set, was greater than the accuracy obtained using conventional fuzzy artificial neural network. Here, the accuracy of algorithm used was improved due to use of dynamic schema and linear scaling.

Anjan Nikhil Repaka, et al., proposed a system in [4] that uses NB (Naïve Bayesian) techniques for classification of



dataset and AES (Advanced Encryption Standard) algorithm for secure data transfer for prediction of disease.

Theresa Princy. R, et al, executed a survey including different classification algorithm used for predicting heart disease. The classification techniques used were Naive Bayes, KNN (K- Nearest Neighbour), Decision tree, Neural network and accuracy of the classifiers was analyzed for different number of attributes [5].

Avinash Golande and et. al.; studies various different ML algorithms that can be used for classification of heart disease. Research was carried out to study Decision Tree, KNN and K-Means algorithms that can be used for classification and their accuracy were compared[1]. This research concludes that accuracy obtained by Decision Tree was highest further it was inferred that it can be made efficient by combination of different techniques and parameter tuning.

The main idea behind the proposed system after reviewing the above papers was to create a heart disease prediction system based on the inputs as shown in Table 1. We analyzed the classification algorithms namely Decision Tree, Random Forest, Logistic Regression and Knn, based on their Accuracy, Precision, Recall and f-measure scores and identified the best classification algorithm which can be used in the heart disease prediction.

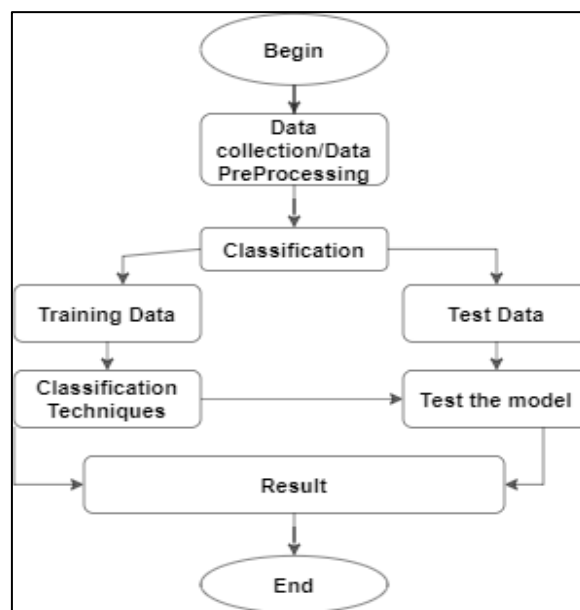


Fig. 1: Generic Model Predicting Heart Disease

#### A. Data Collection and Preprocessing

The dataset used was the Heart disease Dataset which is a combination of 4 different databases, but only the UCI Cleveland dataset was used. This database consists of a total of 76 attributes but all published experiments refer to using a subset of only 14 features [9]. Therefore, we have used the already processed UCI Cleveland dataset available in the Kaggle website for our analysis. The complete description of the 14 attributes used in the proposed work is mentioned in Table 1 shown below.



TABLE I. FEATURES SELECTED FROM DATASET

Sl. No.	Attribute Description	Distinct Values of Attribute
1.	Age- represent the age of a person	Multiple values between 29 & 71
2.	Sex- describe the gender of person (0- Female, 1-Male)	0,1
3.	CP- represents the severity of chest pain patient is suffering.	0,1,2,3
4.	Rest BP-It represents the patient's BP.	Multiple values between 94& 200
5.	Chol-It shows the cholesterol level of the patient.	Multiple values between 126 & 564
6.	FBS-It represent the fasting blood sugar in the patient.	0,1
7.	Resting ECG-It shows the result of ECG	0,1,2
8.	Heartbeat- shows the max heart beat of patient	Multiple values from 71 to 202
9.	Exang- used to identify if there is an exercise induced angina. If yes=1 or else no=0	0,1
10.	Old Peak- describes patient's depression level.	Multiple values between 0 to 6.2
11.	Slope- describes patient condition during peak exercise. It is divided into three segments(Unsloping, Flat, Down sloping)	1,2,3
12.	CA- Result of fluoroscopy.	0,1,2,3
13.	Thal- test required for patient suffering from pain in chest or difficulty in breathing. There are 4 kinds of values which represent Thallium test.	0,1,2,3
14.	Target-It is the final column of the dataset. It is class or label Colum. It represents the number of classes in dataset. This dataset has binary classification i.e. two classes (0,1).In class "0" represent there is less possibility of heart disease whereas "1" represent high chances of heart disease. The value "0" Or "1" depends on other 13 attribute.	0,1

### III. PROPOSED MODEL

The proposed work predicts heart disease by exploring the above mentioned four classification algorithms and does performance analysis. The objective of this study is to effectively predict if the patient suffers from heart disease. The health professional enters the input values from the patient's health report. The data is fed into model which predicts the probability of having heart disease. Fig. 1 shows the entire process involved corresponding branch is followed to that value and jump is made to the next node.

### IV. Logistic Regression

Logistic Regression is a classification algorithm mostly used for binary classification problems. In logistic regression instead of fitting a straight line or hyper plane, the logistic regression algorithm uses the logistic function to squeeze the output of a linear equation between 0 and 1. There are 13 independent variables which makes logistic regression good for classification.

#### A. Classification

The attributes mentioned in Table 1 are provided as input to

### IV. RESULT AND ANALYSIS

(1)

the different ML algorithms such as Random Forest, Decision Tree, Knn, Logistic Regression and The input dataset is split into 70% of the training dataset and the remaining 30% into the test dataset. Training dataset is the dataset which is used to train a model. Testing dataset is used to check the performance of the trained model. For each of the algorithms the performance is computed and analyzed based on different metrics used such as accuracy, precision, and recall and F-measure scores as described further. The different algorithms explored in this paper are listed as below.

#### a. Random Forest

Random Forest algorithms are used for classification as well as regression. It creates a tree for the data and makes prediction based on that. Random Forest algorithm can be used on large datasets and can produce the same result even when large sets record values are missing. The generated samples from the decision tree can be saved so that it can be used on other data. In random forest there are two stages, firstly create a random forest then make a prediction using a



random forest classifier created in the first stage.

#### b. Decision Tree

Decision Tree algorithm is in the form of a flowchart where the inner node represents the dataset attributes and the outer branches are the outcome. Decision Tree is chosen because they are fast, reliable, easy to interpret and very little data preparation is required. In Decision Tree, the prediction of class label originates from root of the tree. The value of the root attribute is compared to record's attribute. On the result of comparison,

The results obtained by applying Random Forest, Decision Tree, Naive Bayes and Logistic Regression are shown in this section. The metrics used to carry out performance analysis of the algorithm are Accuracy score, Precision (P), Recall (R) and F-measure. Precision (mentioned in equation (2)) metric provides the measure of positive analysis that is correct. Recall [mentioned in equation (3)] defines the measure of actual positives that are correct. F-measure [mentioned in equation (4)] tests accuracy.

$$\text{Precision} = (\text{TP}) / (\text{TP} + \text{FP}) \quad (2)$$

$$\text{Recall} = (\text{TP}) / (\text{TP} + \text{FN}) \quad (3)$$

$$\text{F-Measure} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

- TP True positive: the patient has the disease and the test is positive.
- FP False positive: the patient does not have the disease but the test is positive.
- TN True negative: the patient does not have the disease and the test is negative.
- FN False negative: the patient has the disease but the test is negative.

In the experiment the pre-processed dataset is used to carry out the experiments and the above mentioned algorithms are explored and applied. The above mentioned performance metrics are obtained using the confusion matrix. Confusion Matrix describes the performance of the model. The confusion matrix obtained by the proposed model for different algorithms is shown below in Table 2. The accuracy score obtained for Random Forest, Decision Tree, and Logistic Regression and is shown below in Table 3.

TABLE II. VALUES OBTAINED FOR CONFUSION MATRIX USING DIFFERENT ALGORITHM

Algorithm	True Positive	False Positive	False Negative	True Negative
Logistic Regression	22	5	4	30
RandomForest	22	5	6	28
Decision Tree	25	2	4	30

provide better results and help health professionals in predicting the heart disease effectively and efficiently.

#### iii Knn

It is one of the simplest supervised classification algorithm and its most used algorithms ..it is used to solve both classification and regression problems it's easy to understand and implement .

#### Approach and Methodolog

TABLE III. ANALYSIS OF MACHINE LEARNING ALGORITHM

Algorithm	Precision	Recall	F-measure	Accuracy
DecisionTree	0.845	0.823	0.835	81.97%
Logistic Regression	0.857	0.882	0.869	85.25%
RandomForest	0.937	0.882	0.909	90.16%

## V. CONCLUSION

With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of Decision Tree, Logistic Regression, Random Forest and Knn for predicting heart disease using UCI machine learning repository dataset. The result of this study indicates that the Logistic Regression is the most efficient algorithm with accuracy score of 90.16% for prediction of heart disease. In future the work can be enhanced by developing a web application based on the Logistic Regression as



well as using a larger dataset as compared to the one used in this analysis which will help to

Introduction World Health Organization has estimated 12 million deaths occurred worldwide every year due to heart disease. Half deaths in India and other developed countries due to cardiovascular diseases can aid in making decision on lifestyles changes in high risk patients and in turn reduce the complications .

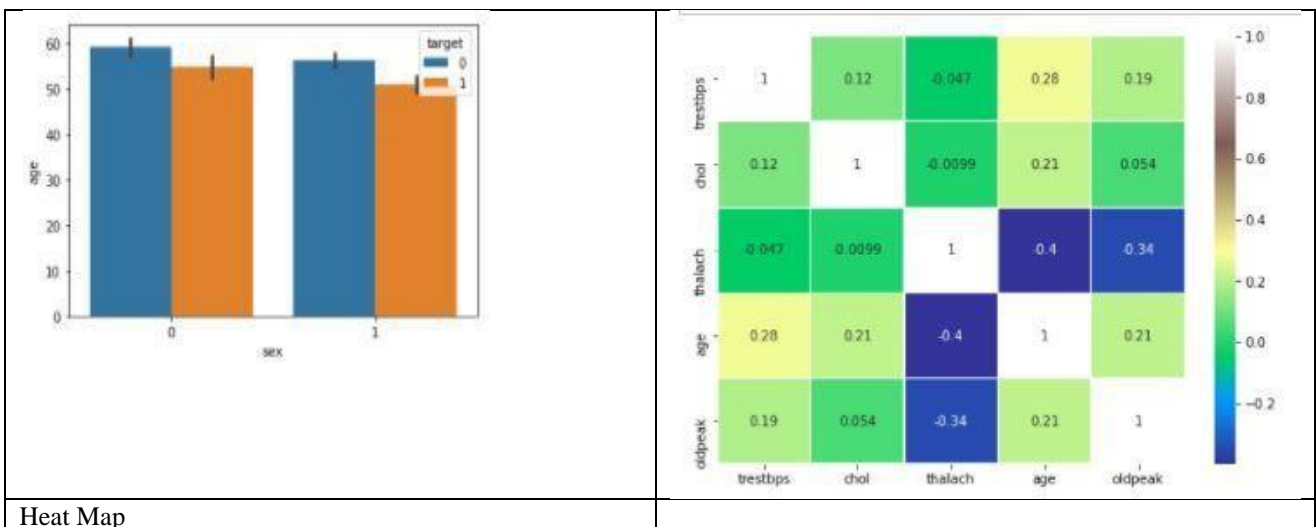
With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of Decision Tree, Logistic Regression, Random Forest and Knn for predicting heart disease using UCI machine learning repository dataset. The result of this study indicates that the Logistic Regression is the most efficient algorithm with accuracy score of 90.16% for prediction of heart disease. In future the work can be enhanced by developing a web application based on the Logistic Regression as well as using a larger dataset as compared to the one used in this analysis which will help to

Introduction World Health Organization has estimated 12 million deaths occurred worldwide every year due to heart disease. Half deaths in India and other developed countries dueto cardiovascular diseases can aid in making decision on lifestyles changes in high risk patients and in turn reduce the complications .



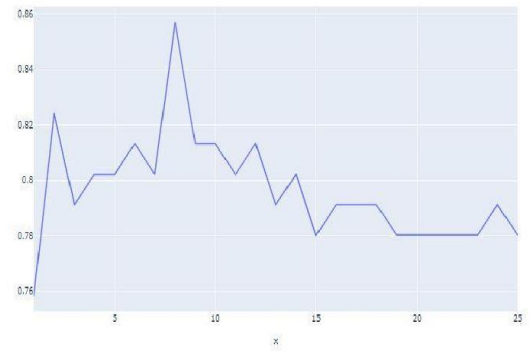
0 NO HEART DISEASE FOR WOMEN WITH 58% AND 48% WOMEN HAVE HEART DISEASE

0 NO HEART DISEASE FOR MEN WITH 55% AND 45% MEN HAVE HEART DISEASE

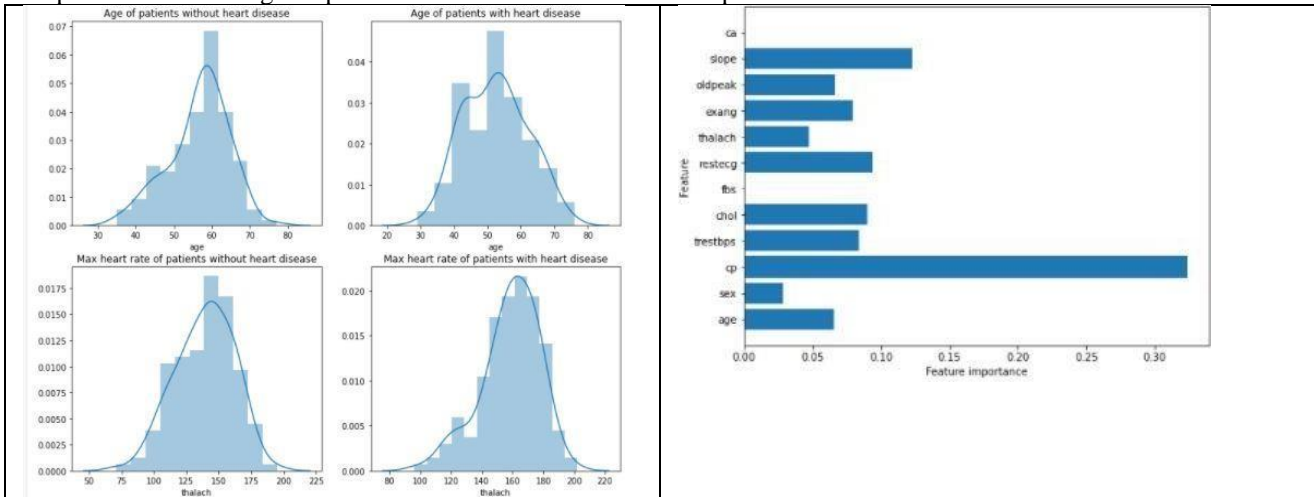




With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of Decision Tree, Logistic Regression, Random Forest and Naive Bayes algorithms for predicting heart disease using UCI machine learning repository dataset. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of 90.16% for prediction of heart disease. In future the work can be enhanced by developing a web application based on the Random Forest algorithm as well as using a larger dataset as compared to the one used in this analysis.



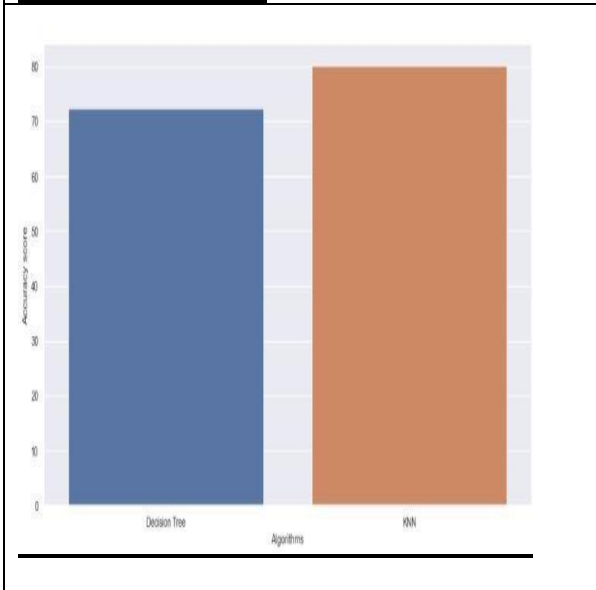
Displots that shows Age of patients with heart disease andno of patients who don't have heart disease



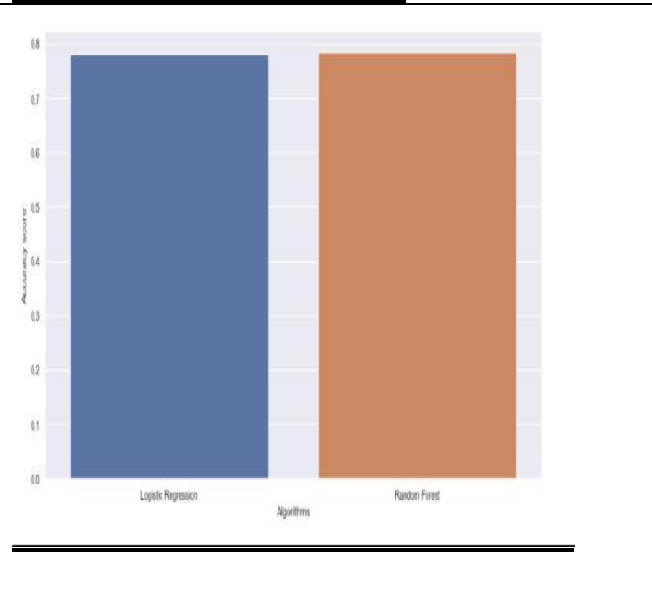
Feature Selection

Accuracy Level Of Future Selection

Final Result Of Accuracy Compared between Decision Tree and Knn



Final Result Of Accuracy Compared between Decision Tree and Knn and logistic regression







## EXISTING SYSTEM

Heart Disease is even highlighted as a silent killer which leads to death of the person without obvious symptoms. The before all existing system works on sets of both Deep learning and Data Mining. Medical Diagnosis plays a vital role and yet complicated task that needs to be executed efficiently and accurately. To reduce the cost for achieving clinical tests appropriate computed based information and decision support should be aided.. Data Mining is the use of software techniques for finding patterns and consistency insets of data .also with the advent of data mining in the last two decades, there is a big opportunity to allow computers to directly construct and classify the different attributes of classes Learning of the risk components connected with heart disease helps medicinal services experts to recognize patients at high risk of having heart disease statistical analysis has identified risk factors associated with heart disease.

## PROPOSED SYSTEM

This proposed system has data which classified if patients have heart disease or not according to features in it. This proposal system can try to use this data to create a model that tries to predict if a patient has a disease or not. In this proposed system using a Decision tree, Random Forest, Logistic Regression, Knn algorithm By using a Skit Library to calculate the score. Random search is a technique where random combinations of the hyperparameters are used to find the best solution for the build model. Finally analyzing the results with the help of comparing models. From the data we are having it should be classified into different structured data based on the features of the patient heart. From the availability of the data. We have to create a model that predicts the patient's disease using a Decision tree, Random Forest, Logistic Regression, Knn algorithm. First we have to import datasets .Read the datasets that should contain different variables like age, gender ,sex, chest pain, slope target. The data should be explored so that the information is verified. Create a temporary variable and also build a Decision tree, Random Forest, Logistic Regression, Knn algorithm model. Hence we use a sigmoid function which helps in graphical representation of the classified data. By using Random Forest and compared to other algorithms the accuracy is increased as compared to previous work done in existing system.

## ACKNOWLEDGMENT

First and Foremost, We are thankful to the Dr. Ambedhkar Institute Of Technology. Master Of Computer Application Department and Dr. Shailaja k Associate Professor, Dr. Ambedhkar Institute Of Technology. Master Of Computer Application Department .A special word of gratitude to Dr. Manjunath, Head of Department, Dr. Ambedhkar Institute Of Technology. Master Of Computer Application Department, for his continued guidance and support for our project work.

## REFERENCES

- [1] Avinash Golande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering, Vol 8, pp.944-950,2019.
- [2] T.Nagamani, S. Logeswari, B. Gomathy," Heart Disease Prediction using Data Mining with Mapreduce Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.
- [3] Fahd Saleh Alotaibi," Implementation of Machine Learning Model to Predict Heart Failure Disease", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.
- [4] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, "Design And Implementation Heart Disease Prediction Using Naives Bayesian", International Conference on Trends in Electronics and Information(ICOEI 2019).
- [5] Theresa Princy R,J. Thomas, 'Human heart Disease Prediction System using Data Mining Techniques', International Conference on Circuit Power and Computing Technologies,Bangalore,2016.
- [6] Nagaraj M Lutimath, Chethan C. Basavaraj S Pol., 'Prediction Of Heart Disease using Machine Learning', International journal Of Recent Technology and Engineering,8,(2S10), pp 474-477, 2019.
- [7] UCI, —Heart Disease Data Set.[Online]. Available (Accessed on May1 2020): <https://www.kaggle.com/ronitf/heart-disease-uci>.
- [8] Sayali Ambekar, Rashmi Phalnikar, "Disease Risk Prediction by Using Convolutional Neural Network",2018 Fourth International Conference on Computing Communication Control and Automation.
- [9] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, —Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field, I in Machine Learning Paradigms, 2019, pp. 71–99.
- [10] Jafar Alzubi, Anand Nayyar, Akshi Kumar. "Machine Learning from Theory to Algorithms: An Overview", Journal of Physics: Conference Series, 2018
- [11] Fajr Ibrahim Alarsan., and Mamoon Younes 'Analysis and classification of heart diseases using heartbeat features and machine learning algorithms', Journal Of Big Data,2019;6:81.
- [12] Internet source [Online].Available (Accessed on May 1 2020): <http://acadpubl.eu/ap>