2020 International Conference on Electrical and Electronics Engineering (ICE3-2020)

Heart Disease Prediction Using Machine Learning Algorithms

Abstract—**Heart plays significant role in living organisms. Diagnosis and prediction of heart related diseases requires more precision, perfection and correctness because a little mistake can cause fatigue problem or death of the person, there are numerous death cases related to heart and their counting is increasing exponentially day by day. To deal with the problem there is essential need of prediction system for awareness about diseases. Machine learning is the branch of Artificial Intelligence(AI), it provides prestigious support in predicting any kind of event which take training from natural events. In this paper, we calculate accuracy of machine learning algorithms for predicting heart disease, for this algorithms are k-nearest neighbor, decision tree, linear regression and support vector machine(SVM) by using UCI repository dataset for training and testing. For implementation of Python programming Anaconda(jupytor) notebook is best tool, which have many type of library, header file, that make the work more accurate and precise.**

Keywords—supervised; unsupervised; reinforced; linear regression; decision tree; python programming; jupytor Notebook; confusion matrix;

I. Introduction

Heart is one of the most extensive and vital organ of human body so the care of heart is essential. Most of diseases are related to heart so the prediction about heart diseases is necessary and for this purpose comparative study needed in this field, today most of patient are died because their diseases are recognized at last stage due to lack of accuracy of instrument so there is need to know about the more efficient algorithms for diseases prediction.

Machine Learning is one of the efficient technology for the testing, which is based on training and testing. It is the branch of Artificial Intelligence(AI) which is one of broad area of learning where machines emulating human abilities, machine learning is a specific branch of AI. On the other hand machines learning systems are trained to learn how to process and make use of data hence the combination of both technology is also called as Machine Intelligence.

As the definition of machine learning, it learns from the natural phenomenon, natural things so in this project we uses the biological parameter as testing data such as cholesterol, Blood pressure, sex, age, etc. and on the basis of these, comparison is done in the terms of accuracy of algorithms such as in this project we have used four algorithms which are decision tree, linear regression, k-neighbour, SVM.

In this paper, we calculate the accuracy of four different machine learning approaches and on the basis of calculation we conclude that which one is best among them.

Section 1 of this paper consist the introduction about the machine learning and heart diseases. Section II described, the machine learning classification. Section III illustrated the related work of researchers. Section IV is about the methodology used for this prediction system. Section V is about the algorithms used in this project. Section VI briefly describes the dataset and their analysis with the result of this project. And the last Section VII concludes the summary of this paper with slight view about future scope of this paper.

II. MACHINE LEARNING

Machine Learning is one of efficient technology which is based on two terms namely testing and training i.e. system take training directly from data and experience and based on this training test should be applied on different type of need as per the algorithm required.

There are three type of machine learning algorithms:

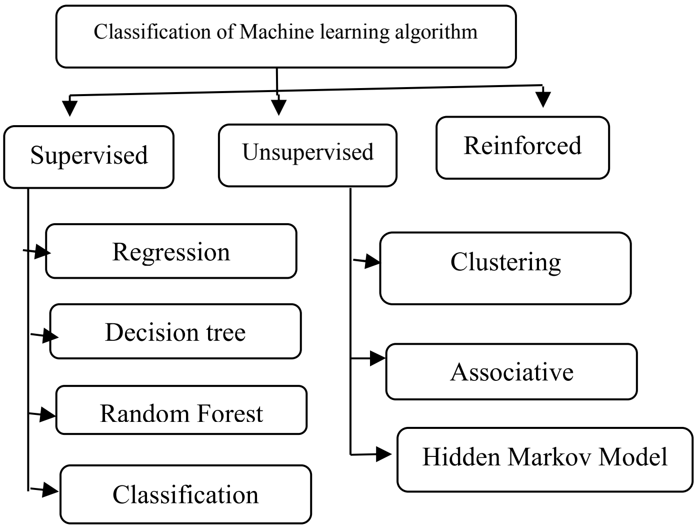


Fig.1 Classification of machine learning

978-1-7281-5846-4/20/$31.00 ©2020 IEEE

A. *Supervised Learning*

Supervised learning can be define as learning with the proper guide or you can say that learning in the present of teacher .we have a training dataset which act as the teacher for prediction on the given dataset that is for testing a data there are always a training dataset. Supervised learning is based on "train me" concept. Supervised learning have following processes:

- Classification
- Random Forest
- Decision tree
- Regression

To recognize patterns and measures probability of uninterruptable outcomes, is phenomenon of regression. System have ability to identify numbers, their values and grouping sense of numbers which means width and height, etc. There are following supervised machine learning algorithms:

- Linear Regression
- Logistical Regression
- Support Vector Machines (SVM)
- Neural Networks
- Random Forest
- Gradient Boosted Trees
- Decision Trees
- Naive Bayes

B. *Unsupervised Learning*

Unsupervised learning can be define as the learning without a guidance which in Unsupervised learning there are no teacher are guiding. In Unsupervised learning when a dataset is given it automatically work on the dataset and find the pattern and relationship between them and according to the created relationships, when new data is given it classify them and store in one of them relation . Unsupervised learning is based on "self sufficient " concept.

For example suppose there are combination fruits mango, banana and apple and when Unsupervised learning is applied it classify them in three different clusters on the basis if there relation with each other and when a new data is given it automatically send it to one of the cluster .

Supervisor learning say there are mango, banana and apple but Unsupervised learning said it as there are three different clusters. Unsupervised algorithms have following process:

- Dimensionality
- Clustering

There are following unsupervised machine learning algorithms:

- t-SNE
- k-means clustering
- PCA

C. *Reinforcement*

Reinforced learning is the agent ability to interact with the environment and find out the outcome. It is based on "hit and trial" concept. In reinforced learning each agent is awarded with positive and negative points and on the basis of positive points reinforced learning give the dataset output that is on the basis of positive awards it trained and on the basis of this training perform the testing on datasets

III. RELATED WORK

Heart is one of the core organ of human body, it play crucial role on blood pumping in human body which is as essential as the oxygen for human body so there is always need of protection of it, this is one of the big reasons for the researchers to work on this. So there are number of researchers working on it .There is always need of analysis of heart related things either diagnosis or prediction or you can say that protection of heart disease .There are various fields like artificial intelligence, machine learning, data mining that contributed on this work .

Performance of any algorithms depends on variance and biasness of dataset[4]. As per research on the machine learning for prediction of heart diseases himanshu et al.[4] naive bayes perform well with low variance and high biasness as compare to high variance and low biasness which is knn. With low biasness and high variance knn suffers from the problem of over fitting this is the reason why performance of knn get decreased. There are various advantage of using low variance and high biasness because as the dataset small it take less time for training as well as testing od algorithm but there also some disadvantages of using small size of dataset. When the dataset size get increasing the asymptotic errors are get introduced and low biasness, low variance based algorithms play well in this type of cases. Decision tree is one of the non-parametric machine learning algorithm but as we know it suffers from the problem over fitting but it cloud be solve by some over fitting removable techniques. Support vector machine is algebraic and statics background algorithm, it construct a linear separable n-dimensional hyper plan for the classification of datasets.

The nature of heart is complex, there is need of carefully handling of it otherwise it cause death of the person. The severity of heart diseases is classified based on various methods like knn, decision tree, generic algorithm and naïve bayes [3]. Mohan et al.[3] define how you can combine two different approaches to make a single approach called hybrid approach which have the accuracy 88.4% which is more than of all other.

Some of the researchers have worked on data mining for the prediction of heart diseases. Kaur et al.[6] have worked on this and define how the interesting pattern and knowledge are derived from the large dataset. They perform accuracy comparison on various machine learning and data mining

978-1-7281-5846-4/20/$31.00 ©2020 IEEE

approaches for finding which one is best among then and get the result on the favor of svm.

Kumar et al.[5] have worked on various machine learning and data mining algorithms and analysis of these algorithms are trained by UCI machine learning dataset which have 303 samples with 14 input feature and found svm is best among them, here other different algorithms are naivy bayes, knn and decision tree.

Gavhane et al.[1] have worked on the multi layer perceptron model for the prediction of heart diseases in human being and the accuracy of the algorithm using CAD technology. If the number of person using the prediction system for their diseases prediction then the awareness about the diseases is also going to increases and it make reduction in the death rate of heart patient.

Some researchers have work on one or two algorithm for predication diseases. Krishnan et al.[2] proved that decision tree is more accurate as compare to the naïve bayes classification algorithm in their project.

Machine learning algorithms are used for various type of diseases predication and many of the researchers have work on this like Kohali et al.[7] work on heart diseases prediction using logistic regression, diabetes prediction using support vector machine, breast cancer prediction using Adaboot classifier and concluded that the logistic regression give the accuracy of 87.1%, support vector machine give the accuracy of 85.71%, Adaboot classifier give the accuracy up to 98.57% which good for predication point of view.

A survey paper on heart diseases predication have proven that the old machine learning algorithms does not perform good accuracy for the predication while hybridization perform good and give better accuracy for the predication[8].

IV. METHODOLOGY OF SYSTEM

Processing of system start with the data collection for this we uses the UCI repository dataset which is well verified by number of researchers and authority of the UCI [15].

A. Data Collection

First step for predication system is data collection and deciding about the training and testing dataset. In this project we have used 73% training dataset and 37% dataset used as testing dataset the system.

B. Attribute Selection

Attribute of dataset are property of dataset which are used for system and for heart many attributes are like heart bit rate of person, gender of the person, age of the person and many more shown in TABLE.1 for predication system.

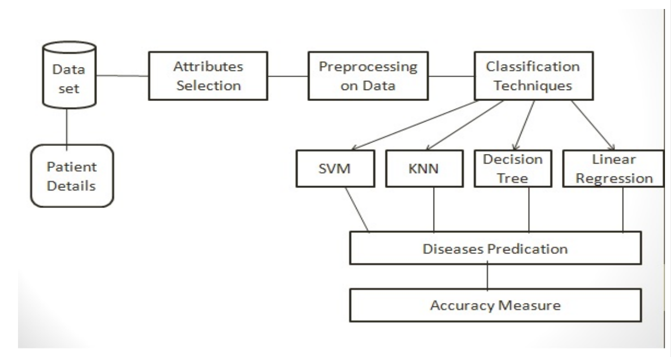


Fig.2 Architecture of Prediction System

TABLE.1 Attributes of the Dataset

S. No.	Attribute	Description	Type
1	Age	Patient's age (29 to 77)	Numaric
2	Sex	Gender of patient(male-0 female-1)	Nominal
3	Cp	Chest pain type	Nominal
4	Trestbps	Resting blood pressure(in mm Hg on admission to hospital ,values from 94 to 200)	Numerical
5	Chol	Serum cholesterol in mg/dl, values from 126 to 564)	Numerical
6	Fbs	Fasting blood sugar>120 mg/dl, true-1 false-0)	Nominal
7	Resting	Resting electrocardiographics result (0 to 1)	Nominal
8	Thali	Maximum heart rate achieved(71 to 202)	Numerical
9	Exang	Exercise included agina(1-yes 0-no)	Nominal
10	Oldpeak	ST depression introduced by exercise relative to rest (0 to .2)	Numerical
11	Slope	The slop of the peak exercise ST segment (0 to 1)	Nominal
12	Ca	Number of major vessels (0-3)	Numerical
13	Thal	3-normal	Nominal
14	Targets	1 or 0	Nominal

978-1-7281-5846-4/20/$31.00 ©2020 IEEE

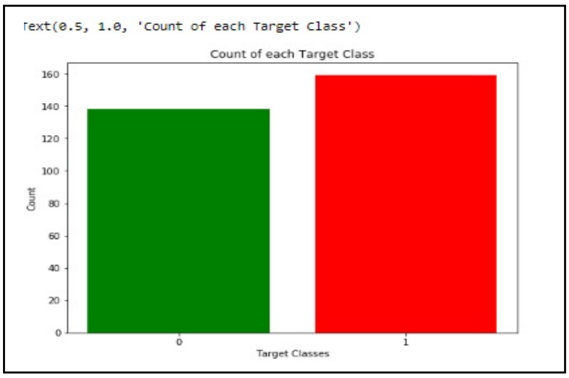
C. *Preprocessing of data*

Preprocessing needed for achieving prestigious result from the machine learning algorithms. For example Random forest algorithm does not support null values dataset and for this we have to manage null values from original raw data.

For our project we have to convert some categorized value by dummy value means in the form of "0"and "1" by using following code:

D. *Data Balancing*

Data balancing is essential for accurate result because by data balancing graph we can see that both the target classes are equal. Fig.3 represents the target classes where "0" represents with heart diseases patient and "1" represents no heart diseases pateints.



```
Text(0.5, 1.0, 'Count of each Target Class')
```

Fig.3 Target class view

E. *Histogram of attributes*

Histogram of attributes shows the range of dataset attributes and code which is used to create it.

dataset.hist()

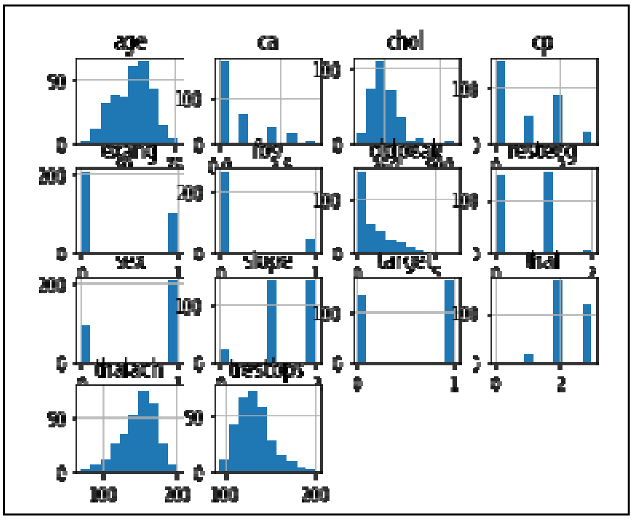


Fig.4 Histogram of attributes

V. MACHINE LEARNING ALGORITHMS

A. *Linear regression*

It is the supervised learning technique. It is based on the relationship between independent variable and dependent variable as seen in Fig.5 variable "x" and "y" are independent and dependent variable and relation between them is shown by equation of line which is linear in nature that why this approach is called linear regression.

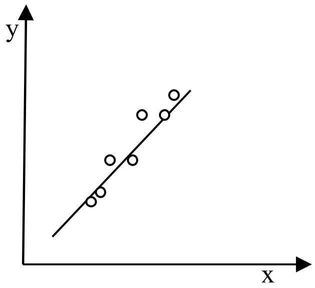


Fig.5 relation between x and y

It gives a relation equation to predict a dependent variable value "y" based on a independent variable value "x" as we can see in the Fig.5 so it is concluded that linear regression technique give the linear relationship between x(input) and y(output).

B. *Decision tree*

On the other hand decision tree is the graphical representation of the data and it is also the kind of supervised machine learning algorithms.

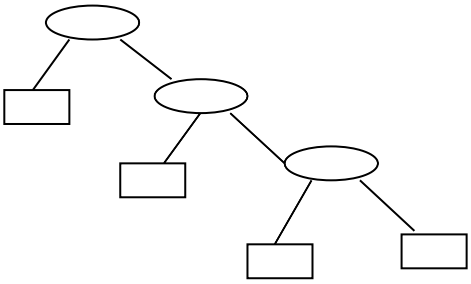


Fig.6 Decision tree

For the tree construction we use entropy of the data attributes and on the basis of attribute root and other nodes are drawn.

$$\text{Entropy} = -\sum P_{ij} \log P_{ij} \qquad (1)$$

In the above equation of entropy (1) Pij is probability of the node and according to it the entropy of each node is calculated. The node which have highest entropy calculation is selected as the root node and this process is repeated until all the nodes of the tree are calculated or until the tree constructed.

When the number of nodes are imbalanced then tree is create the over fitting problem which is not good for the

978-1-7281-5846-4/20/$31.00 ©2020 IEEE

calculation and this is one of reason why decision tree have less accuracy as compare to linear regression.

C. Support Vector Machine

It is one category of machine learning technique which work on the concept of hyperplan means it classify the data by creating hyper plan between them.

Training sample dataset is (Yi, Xi) where i=1,2,3,…….n and Xi is the ith vector, Yi is the target vector. Number of hyper plan decide the type of support vector such as example if a line is used as hyper plan then method is called linear support vector.

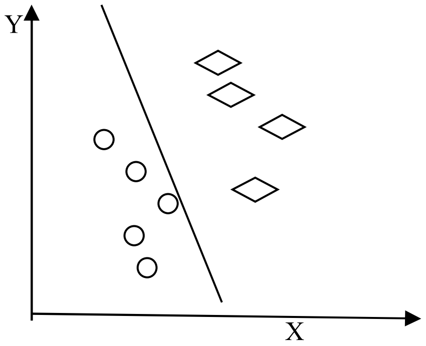


Fig.7 Linear Regression

D. K-nearest Neighbour

It work on the basis of distance between the location of data and on the basis of this distinct data are classified with each other. All the other group of data are called neighbor of each other and number of neighbor are decided by the user which play very crucial role in analysis of the dataset.

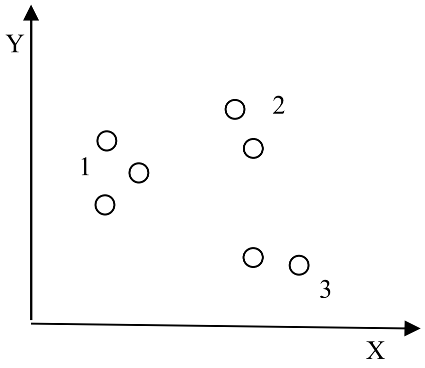


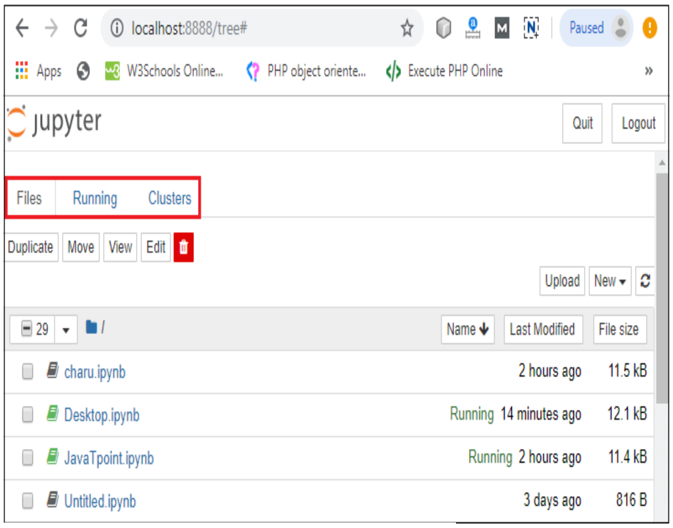
Fig.8 KNN where k=3

In the above Fig. k=3 shows that there are three neighbor that means three different type of data are there. Each cluster represented in two dimensional space whose coordinates are represented as (Xi,Yi) where Xi is the x-axis, Y represent y-axis and i= 1,2,3,….n.

VI. Result Analysis

A. About Jupytor Notebook

Jupiter notebook is used as the simulation tool and it is confortable for python programming projects. Jupytor notebook contains rich text elements and code also, which are figures, equations, links and many more. Because of the mix of rich text elements and code, these documents are perfect location to bring together an analysis description, and its results, as well as, they can execute data analysis in real time. Jupyter Notebook is an open-source, web-based interactive graphics, maps, plots, visualizations, and narrative text.



Name		Last Modified	File size
charu.ipynb		2 hours ago	11.5 kB
Desktop.ipynb	Running	14 minutes ago	12.1 kB
JavaTpoint.ipynb	Running	2 hours ago	11.4 kB
Untitled.ipynb		3 days ago	816 B

Fig.9 Jupyter Notebook

B. Accuracy calculation

Accuracy of the algorithms are depends on four values namely true positive(TP), false positive(FP), true negative(TN) and false negative(FN).

$$\text{Accuracy}= (FN+TP) / (TP+FP+TN+FN) \qquad (2)$$

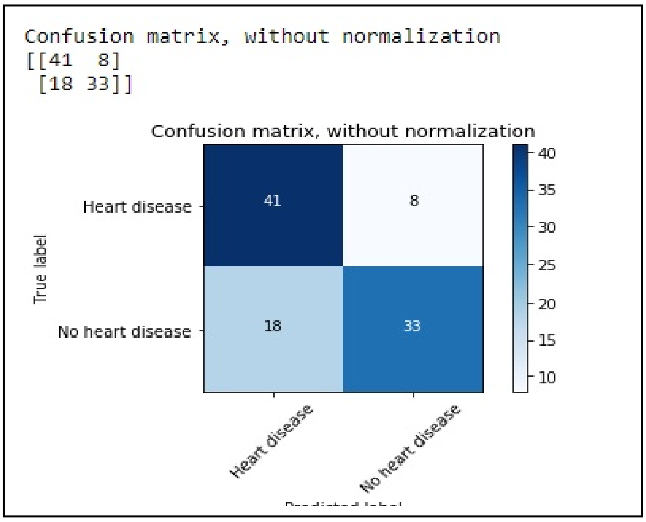
The numerical value of TP, FP, TN, FN defines as:

TP= Number of person with heart diseases

TN= Number of person with heart diseases and no heart diseases

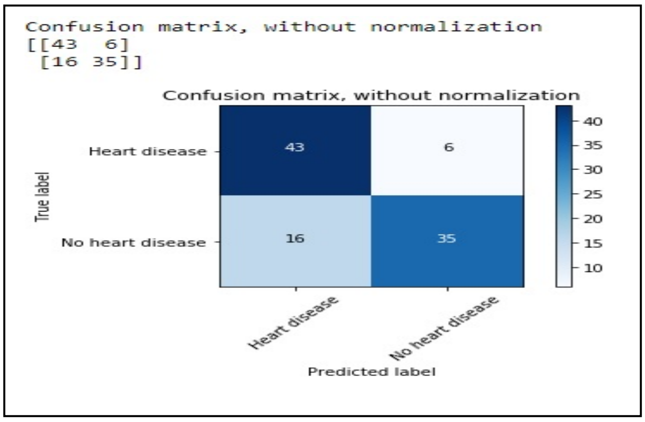
FP= Number of person with no heart diseases

FN= Number of person with no heart diseases and with heart diseases



```
Confusion matrix, without normalization
[[41  8]
 [18 33]]
```

Fig.10 Confusion matrix for Decision tree



```
Confusion matrix, without normalization
[[43  6]
 [16 35]]
```

Fig.11 Confusion Matrix for linear regression

C. Result

After performing the machine learning approach for testing and training we find that accuracy of the knn is much efficient as compare to other algorithms. Accuracy should be calculated with the support of confusion matrix of each algorithms as shown in Fig.6 and Fig.7 here number of count of TP, TN, FP, FN are given and using the equation (2) of accuracy, value has been calculated and it is conclude that knn is best among them with 87% accuracy and the comparison is shown in TABLE.2

TABLE.2 Accuracy comparison

Algorithm	Accuracy
Support Vector machine	83%
Decision tree	79%
Linear regression	78%
k-nearest neighbor	87%

VII. CONCLUSION AND FUTURE SCOPE

Heart is one of the essential and vital organ of human body and prediction about heart diseases is also important concern for the human beings so that the accuracy for algorithm is one of parameter for analysis of performance of algorithms. Accuracy of the algorithms in machine learning depends upon the dataset that used for training and testing purpose. When we perform the analysis of algorithms on the basis of dataset whose attributes are shown in TABLE.1 and on the basis of confusion matrix, we find KNN is best one.

For the Future Scope more machine learning approach will be used for best analysis of the heart diseases and for earlier prediction of diseases so that the rate of the death cases can be minimized by the awareness about the diseases.

References

[1] Santhana Krishnan J and Geetha S, "Prediction of Heart Disease using Machine Learning Algorithms" ICIICT, 2019.

[2] Aditi Gavhane, Gouthami Kokkula, Isha Panday, Prof. Kailash Devadkar, "Prediction of Heart Disease using Machine Learning", Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology(ICECA), 2018.

[3] Senthil kumar mohan, chandrasegar thirumalai and Gautam Srivastva, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" IEEE Access 2019.

[4] Himanshu Sharma and M A Rizvi, "Prediction of Heart Disease using Machine Learning Algorithms: A Survey" International Journal on Recent and Innovation Trends in Computing and Communication Volume: 5 Issue: 8 , IJRITCC August 2017.

[5] M. Nikhil Kumar, K. V. S. Koushik, K. Deepak, "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools" International Journal of Scientific Research in Computer Science, Engineering and Information Technology ,IJSRCSEIT 2019.

[6] Amandeep Kaur and Jyoti Arora,"Heart Diseases Prediction using Data Mining Techniques: A survey" International Journal of Advanced Research in Computer Science , IJARCS 2015-2019.

[7] Pahulpreet Singh Kohli and Shriya Arora, "Application of Machine Learning in Diseases Prediction", 4th International Conference on Computing Communication And Automation(ICCCA), 2018.

[8] M. Akhil, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," Procedia Technol., vol. 10, pp. 85–94, 2013.

[9] S. Kumra, R. Saxena, and S. Mehta, "An Extensive Review on Swarm Robotics," pp. 140–145, 2009.

[10] Hazra, A., Mandal, S., Gupta, A. and Mukherjee, " A Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review" Advances in Computational Sciences and Technology , 2017.

[11] Patel, J., Upadhyay, P. and Patel, "Heart Disease Prediction Using Machine learning and Data Mining Technique" Journals of Computer Science & Electronics , 2016.

[12] Chavan Patil, A.B. and Sonawane, P."To Predict Heart Disease Risk and Medications Using Data Mining Techniques with an IoT Based Monitoring System for Post-Operative Heart Disease Patients" International Journal on Emerging Trends in Technology, 2017.

[13] V. Kirubha and S. M. Priya, "Survey on Data Mining Algorithms in Disease Prediction," vol. 38, no. 3, pp. 124–128, 2016.

[14] M. A. Jabbar, P. Chandra, and B. L. Deekshatulu, "Prediction of risk score for heart disease using associative classification and hybrid feature subset selection," Int. Conf. Intell. Syst. Des. Appl. ISDA, pp. 628–634, 2012.

[15] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

978-1-7281-5846-4/20/$31.00 ©2020 IEEE