DISEASE PREDICTION SYSTEM – SMART HEALTHCARE TOOL using Machine Learning Algorithm

Manisha Mohan Pancholi [1], Prachi Priya[2], Shubham Raina [3], Chakradhar Panchal[4]

[1,2,3,4,] BE Student, Department of Computer Engineering, Sinhgad Institute Of Technology, Lonavala, Maharashtra, India

---***---

Abstract :

The world is moving at a fast speed and to keep up ourselves with the whole world, we tend to ignore the symptoms of disease which can affect our health to a large extent. Many working professionals get a heart attack, bad cholesterol, eye disease and they are unable to treat it at the right time as they are busy coping up with the progressive world. God has granted every individual a beautiful gift called life, so it is our responsibility to live our life to the fullest and try to stay safe from the dangers of the world. So we have developed a logistic regression model for Covid-19 with the machine learning algorithms like decision tree, a random forest which takes into account the symptoms felt by a person, and according to those symptoms it predicts the disease which the person, can be suffering from. It saves time as well as makes it easy to get a warning about your health before it's too late.

***Key Words**: Machine Learning, Kaggle, UCI-REPOSITORY, Decision tree, logistic regression, CKD, COVID-19.*

1. INTRODUCTION

Today, the healthcare industry has become a big money-making business. The healthcare industry uses and produces quite a large amount of data that can be used to extract information about a particular disease for a patient. This information of healthcare will further be used for effective and best possible treatment for patient's health. This area also needs some improvement by using informative data in healthcare sciences. But a major challenge is to extract the information from the data because the data is present in a huge amount so some data mining and machine learning techniques are used. The expected result and of this project is to predict the disease so that the risk of life can be prevented at an early stage and save the life of people and the cost of treatment can be reduced to a particular extent. In India also we should adopt the non-manual system of medical treatment which is the best for improving and understanding human health. The main motive is to use the concept of machine learning in healthcare to improvise the treatment of patients. Machine learning has already made it much easier to identify and predict various diseases. Predictive analysis of the disease with the help of many machine learning algorithms helps us to predict the disease and helps in effectively treating the patients. Disease prediction using machine learning also uses the patient history and health data by applying various concepts like data mining and machine learning techniques and also some algorithm. Many works have also applied data mining techniques to the pathological data for the prediction of some particular diseases. These approaches were intended to beforehand predict the reoccurrence of certain diseases. Also, some approaches tried to do prediction while controlling the disease. The recent work of deep learning was in disparate areas of machine learning which have driven a shift to machine learning models that can learn and understand the hierarchical representations of raw data with some pre-processing. With the development of this concept called big data technology, more attention is paid to disease prediction.

2. OBJECTIVES

Nowadays, people face various diseases due to environmental condition and their living habits. So the prediction of disease at an earlier stage becomes an important task. But the accurate prediction based on symptoms becomes too difficult for the doctor. The correct prediction of disease is the most challenging task. To overcome this problem data mining plays an important role to predict the disease. Medical science has a large amount of data growth per year. Due to the increasing amount of data growth in the medical and healthcare field the accurate analysis on medical data has been benefiting from early patient care. With the help of disease data, data mining finds hidden pattern information in a huge amount of medical data. We proposed general disease prediction based on the symptoms of the patient. For the disease prediction, we use Decision tree Classifier for Chronic Kidney Disease and general disease Prediction, Logistic regression for Covid-19 prediction machine learning algorithm for accurate prediction of disease, Random forest for Cardiovascular disease prediction. For disease prediction required disease symptoms dataset. In this general disease prediction, the living habits of a person and checkup information consider for the accurate prediction. The accuracy of general disease prediction by using a Decision tree is 97% which is more than the

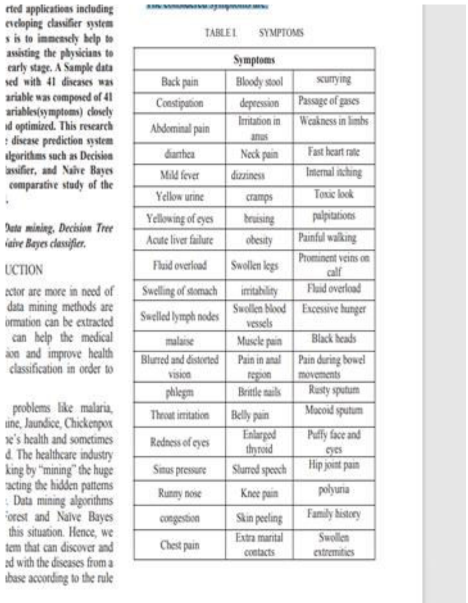
Logistic regression algorithm. Similarly, accuracy is the measure for each disease using a confusion matrix.

2. OVERVIEW

The dataset we have considered consists of 132 symptoms, the combination or permutations of which leads to 41 diseases. Based on the 4920 records of patients, we aim to develop a prediction model that takes in the symptoms from the user and predicts the disease he is more likely to have. We created 4 models in our system:

- Chronic Kidney Disease Predictor
- Corona Virus (COVID-19) Predictor
- Cardiovascular Disease (CVD) Predictor
- General Disease Predictor

The considered symptoms are:



rted applications including
eveloping classifier system
s is to immensely help to
assisting the physicians to
early stage. A Sample data
sed with 41 diseases was
ariable was composed of 41
ariables(symptoms) closely
d optimized. This research
: disease prediction system
lgorithms such as Decision
lassifier, and Naïve Bayes
comparative study of the
.

Data mining, Decision Tree
aive Bayes classifier.

UCTION

ector are more in need of
data mining methods are
ormation can be extracted
can help the medical
ion and improve health
classification in order to

problems like malaria,
ine, Jaundice, Chickenpox
e's health and sometimes
d. The healthcare industry
king by "mining" the huge
acting the hidden patterns
. Data mining algorithms
orest and Naïve Bayes
this situation. Hence, we
tem that can discover and
ed with the diseases from a
base according to the rule

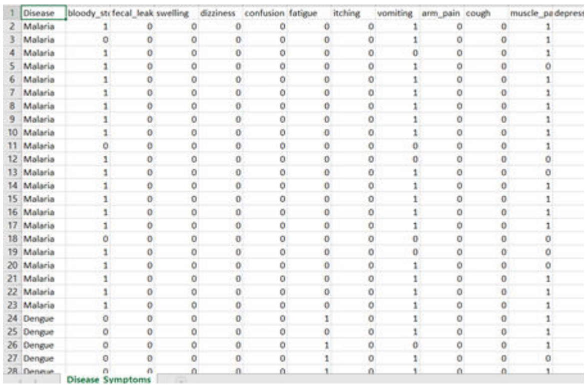
TABLE I. SYMPTOMS

Symptoms		
Back pain	Bloody stool	scurrying
Constipation	depression	Passage of gases
Abdominal pain	Irritation in anus	Weakness in limbs
diarrhea	Neck pain	Fast heart rate
Mild fever	dizziness	Internal itching
Yellow urine	cramps	Toxic look
Yellowing of eyes	bruising	palpitations
Acute liver failure	obesity	Painful walking
Fluid overload	Swollen legs	Prominent veins on calf
Swelling of stomach	irritability	Fluid overload
Swelled lymph nodes	Swollen blood vessels	Excessive hunger
malaise	Muscle pain	Black heads
Blurred and distorted vision	Pain in anal region	Pain during bowel movements
phlegm	Brittle nails	Rusty sputum
Throat irritation	Belly pain	Mucoid sputum
Redness of eyes	Enlarged thyroid	Puffy face and eyes
Sinus pressure	Slurred speech	Hip joint pain
Runny nose	Knee pain	polyuria
congestion	Skin peeling	Family history
Chest pain	Extra marital contacts	Swollen extremities

Fig. 1. Attributes Used in General Disease Prediction System.

2.1 General Disease Predictor:

Dataset has taken for general disease prediction system from UCI-Repository which contain records of more than 5000 patients. The accuracy of 97% is archive through the Decision Tree algorithm.



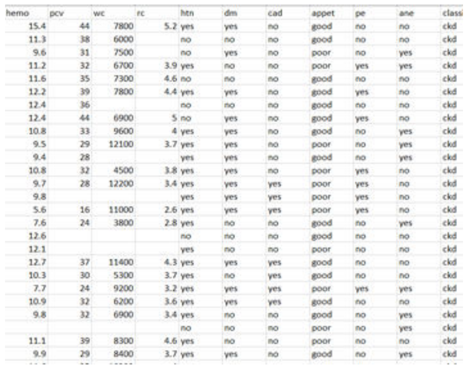
1	Disease	bloody_st	fecal_leak	swelling	dizziness	confusion	fatigue	itching	vomiting	arm_pain	cough	muscle_pa	depres
2	Malaria	1	0	0	0	0	0	0	1	0	0	1	
3	Malaria	0	0	0	0	0	0	0	1	0	0	1	
4	Malaria	1	0	0	0	0	0	0	0	0	0	1	
5	Malaria	1	0	0	0	0	0	0	1	0	0	0	
6	Malaria	1	0	0	0	0	0	0	1	0	0	1	
7	Malaria	1	0	0	0	0	0	0	1	0	0	1	
8	Malaria	1	0	0	0	0	0	0	1	0	0	1	
9	Malaria	1	0	0	0	0	0	0	1	0	0	1	
10	Malaria	1	0	0	0	0	0	0	1	0	0	1	
11	Malaria	0	0	0	0	0	0	0	0	0	0	1	
12	Malaria	1	0	0	0	0	0	0	0	0	0	0	
13	Malaria	1	0	0	0	0	0	0	1	0	0	0	
14	Malaria	1	0	0	0	0	0	0	1	0	0	1	
15	Malaria	1	0	0	0	0	0	0	1	0	0	1	
16	Malaria	1	0	0	0	0	0	0	1	0	0	1	
17	Malaria	1	0	0	0	0	0	0	1	0	0	1	
18	Malaria	0	0	0	0	0	0	0	0	0	0	0	
19	Malaria	1	0	0	0	0	0	0	0	0	0	0	
20	Malaria	1	0	0	0	0	0	0	1	0	0	0	
21	Malaria	1	0	0	0	0	0	0	1	0	0	1	
22	Malaria	1	0	0	0	0	0	0	1	0	0	1	
23	Malaria	1	0	0	0	0	0	0	1	0	0	1	
24	Dengue	0	0	0	0	0	1	0	1	0	0	1	
25	Dengue	0	0	0	0	0	0	0	1	0	0	1	
26	Dengue	0	0	0	0	0	1	0	0	0	0	1	
27	Dengue	0	0	0	0	0	1	0	1	0	0	0	
28	Dengue	0	0	0	0	0	1	0	1	0	0	1	

Disease Symptoms

Fig . 2 . Dataset from UCI-Repository for General Disease Prediction

2.2 Chronic Kidney Disease (CKD) Predictor:

Dataset used for kidney disease prediction from kaggle.com which contains records of more than 400 patients. The Algorithm used to predict kidney disease is the Decision tree it provides 100% accuracy. Attributes choose for this model are sugar christening, Albumin, Hypertension, Packed volume cell, Serum christening, Specific gravity, etc.



hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classi
15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
11.3	38	6000		no	no	no	good	no	no	ckd
9.6	31	7500		no	yes	no	poor	no	yes	ckd
11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
11.6	35	7300	4.6	no	no	no	good	no	no	ckd
12.2	39	7800	4.4	yes	yes	no	good	yes	no	ckd
12.4	36			no	no	no	good	no	no	ckd
12.4	44	6900	5	no	yes	no	good	yes	no	ckd
10.8	33	9600	4	yes	yes	no	good	no	yes	ckd
9.5	29	12100	3.7	yes	yes	no	poor	no	yes	ckd
9.4	28			yes	yes	no	good	no	yes	ckd
10.8	32	4500	3.8	yes	yes	no	poor	yes	no	ckd
9.7	28	12200	3.4	yes	yes	yes	poor	yes	no	ckd
9.8				yes	yes	yes	poor	yes	no	ckd
5.6	16	11000	2.6	yes	yes	yes	poor	yes	no	ckd
7.6	24	3800	2.8	yes	no	no	good	no	yes	ckd
12.6				no	no	no	good	no	no	ckd
12.1				yes	no	no	poor	no	no	ckd
12.7	37	11400	4.3	yes	yes	yes	good	no	no	ckd
10.3	30	5300	3.7	yes	no	yes	good	no	no	ckd
7.7	24	9200	3.2	yes	yes	yes	poor	yes	yes	ckd
10.9	32	6200	3.6	yes	yes	yes	good	no	no	ckd
9.8	32	6900	3.4	yes	no	no	good	no	yes	ckd
				no	no	no	poor	no	yes	ckd
11.1	39	8300	4.6	yes	no	no	poor	no	no	ckd
9.9	29	8400	3.7	yes	yes	no	good	no	yes	ckd

Fig . 3. Dataset from kaggle for prediction of Kidney Disease

2.3 Cardiovascular Disease (CVD) Predictor:

Heart Disease is the most common disease in today's era, There are lots of people who die every year due to these diseases. To overcome the death rate we predicted this system with nearly 98% accuracy result, we use here a Random Forest Machine learning algorithm. Dataset chooses from Kaggle.

If age=<30 and Overweight=no and Alcohol Intake=never. Then

Heart Disease level is Low.
(OR)
If Age=> and Blood Pressure=High and smoking=current
Then
Heart Disease level is High.

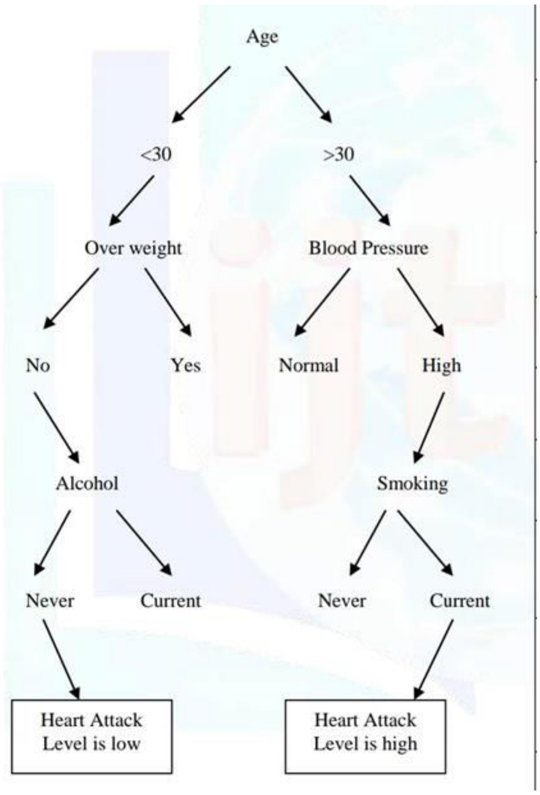
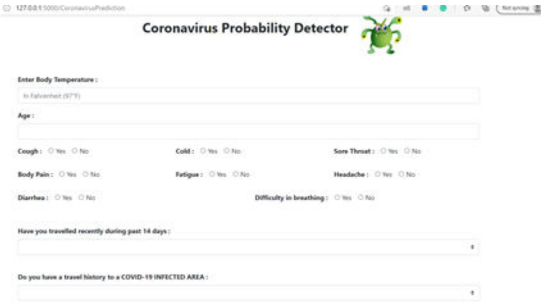


Fig. 4. Heart Disease Prediction Using Random Forest.

2.4 Corona Virus (COVID-19) Predictor:

As we all know in the current pandemic situation it's too risky to go outside. So, we created this model with 96% accuracy to predict COVID-19 diseases at one touch. The attribute used in these models are breathing, Cough, Cold, Fever, Oxygen level, travelled history etc. Logistic regression algorithm gives the best result for this model



Coronavirus Probability Detector

Enter Body Temperature :
In Fahrenheit (97°F)

Age :

Cough : ○ Yes ○ No Cold : ○ Yes ○ No Sore Throat : ○ Yes ○ No

Body Pain : ○ Yes ○ No Fatigue : ○ Yes ○ No Headache : ○ Yes ○ No

Diarrhea : ○ Yes ○ No Difficulty in breathing : ○ Yes ○ No

Have you travelled recently during past 14 days :

Do you have a travel history to a COVID-19 INFECTED AREA :

Fig. 5. Covid-19 Predictor

3. METHODOLOGY

Disease prediction system follows following steps:

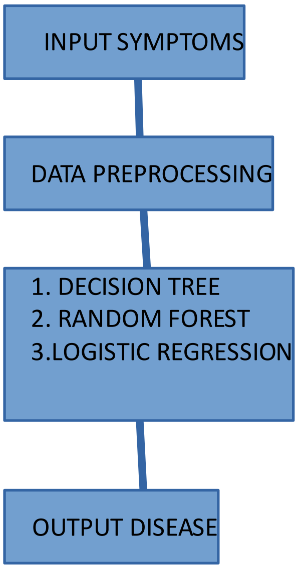


Fig. 6. PREDICTION MODEL

A. Input (Symptoms):

While designing the model we have assumed that the user has a clear idea about the symptoms he is experiencing. The Prediction developed considers 95 symptoms amidst which the user can give the symptoms his processing as the input.

B. Data preprocessing:

The data mining technique that transforms the raw data or encodes the data to a form which can be easily interpreted by the algorithm is called data preprocessing. The preprocessing techniques used in the presented work are:

- Data Cleaning: Data is cleansed through processes such as filling in missing value, thus resolving the inconsistencies in the data.

- Data Reduction: The analysis becomes hard when dealing with huge database. Hence, we eliminate those independent variables(symptoms) which might have less or no impact on the target variable(disease). In the present work, 95 of 132 symptoms closely related to the diseases are selected.

C. Models selected:

The system is trained to predict the diseases using three algorithms x Disease Tree Classifier x Random forest Classifier x Naïve Bayes Classifier A comparative study is presented at the end of work, thus analyzing the performance of each algorithm of the considered database.

D. Output(diseases):

Once the system is trained with the training set using the mentioned algorithms a rule set is formed and when the user the symptoms are given as an input to the model. those symptoms are processed according the rule set developed,

3.1. ALGORITHMS USED:

A. Decision Tree Classifier:

The classification models built by decision tree resemble the structure of tree. By learning the series of explicit if-then rules on feature values (symptoms in our case), it breaks down the dataset into smaller and smaller subsets that results in predicting a target value(disease). A decision tree consists of the decision nodes and leaf nodes.

- Decision node: Has two or more branches. In our work presented, all the symptoms are considered as decision nodes.
- Leaf node: Represents the classification that is, the Decision of any branch. Here the Diseases correspond to the leaf nodes

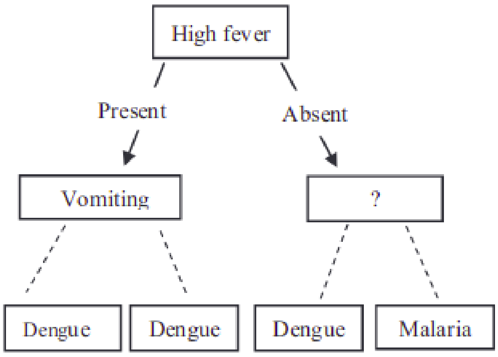


Fig. 7. DECISION TREE FOR GENERAL DISEASE PREDICTION

B. Logistic Regression :

It is a machine learning approach, which can be used for classification problem. In a paper published in Proceedings of Belgian Royal Academy, Pierre Francais Verhulst described the logistic function and its properties by defining three parameters and the curve passing through these.[5]. It is very simple and one of the most used machine learning algorithms. Logistic regression is a statistical model for predicting binary classes. The maximum likelihood Estimation model is used in logistic regression. The dependent variable here follows the Bernoulli distribution. Logistic function or sigmoid function is an 'S' shaped curve that takes a value between 0 and 1. 1 will be predicted if the curve goes to positive infinity and 0 if it goes to negative infinity. Figure 1 is a representation of a logistic function

The above equation is the representation of a logistic regression

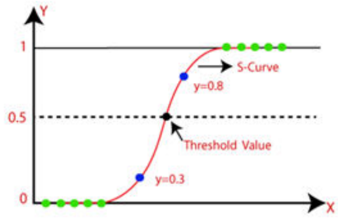


Figure 1: Representation of Logistic Regression Model.[6]

Fig. 8. LOGISTIC REGRESSION MODEL FOR COVID-19

C. Random Forest Algorithm:

Random forest is a flexible, easy to use machine learning algorithm that provides exceptional results most of the time even without hyper-tuning. As mentioned in the Decision tree, the major limitation of the decision tree algorithm is overfitting. It appears as if the tree has memorized the data. Random Forest prevents this problem: It is a version of ensemble learning. Ensemble learning refers to using multiple algorithms or the same algorithm multiple times. Random forest is a team of Decision trees. And greater the number of these decision trees in Random forest, the better the generalization. More precisely, Random forest works as follows:

1. Selects k symptoms from dataset (medical record) with a total of m symptoms randomly (where k<<m) Then, it builds a decision tree from those k symptoms.
2. Repeats n times so that we have n decision trees built from different random combinations of k symptoms (or a different random sample of the data, called bootstrap sample)

3. Takes each of the n-built decision trees and passes a random variable to predict the Disease. Stores the predicted Disease, so that we have a total of n Diseases predicted from n Decision trees.
4. Calculates the votes for each predicted Disease and takes the mode (most frequent Disease predicted) as the final prediction from the random forest algorithm.

4. RESOURCES USED

4.1 Hardware:

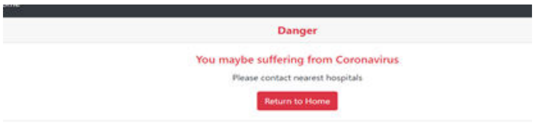
1. Processor: Minimum 2.0GHz requires.
2. Ram: 2 GB.
3. Hard Disk: 100 GB.
4. Input device: Standard Keyboard and Mouse.
5. Output device: VGA and High-Resolution Monitor

4.2 Software:

1. Operating System: Windows 7.
2. Language: Python
3. Algorithm Used: Decision Tree, Random Forest, Logistic regression
4. Tool: Jupyter Notebook, Visual Studio.

5. RESULT AND ANALYSIS

When person having chances to particular diseases it display the following result:

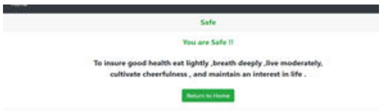

Danger

You maybe suffering from Coronavirus

Please contact nearest hospitals

Fig. 9. Person causes Disease

When person is safe for does not causes any disease it display the following result:


Safe

You are Safe !!

To insure good health eat lightly ,breath deeply ,live moderately, cultivate cheerfulness , and maintain an interest in life .

Fig. 10. Person does not causes Disease or person is safe

The results obtained by applying Random Forest, Decision Tree 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 accuracy.

(1)Precision = (TP) / (TP +FP)

(2) Recall = (TP) / (TP+FN)

(3) F– Measure =(2 * Precision * Recall) / (Precision +Recall)

(4) **Accuracy=TP+TN/TP+FP+FN+TN**

- 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.

Diseases Algorithm	Heart Disease Prediction	General Disease Prediction	Corona Virus Disease Prediction	Kidney Disease Prediction
Logistic Regression	0.696533	-	0.921111	0.94657
Decision Tree	0.99971	0.98154	_	1.00000
Random Forest	0.97568		_	1.00000

Fig. 11. Accuracy Measure using confusion Matrix

7. CONCLUSION AND FUTURE WORK

In the proposed system, hidden knowledge will be extracted from the historical data by preparing datasets by applying the Naïve Byes algorithm. Predicting smart health can be done only the system responds that way. These datasets will be compared with the incoming queries and the final report will be generated using Association Rule

Mining. Since this proposed methodology will work on real historical data, it will provide accurate and efficient results, which will help patients get diagnosis instantly. This system will also guide the users on how to remain healthy and fit using tips provided here. The further enhancements that can be done would be integrating this web application into an Android app. This will be available to users on a mobile basis and its use can be further increased. Also feature like getting the doctor online on chat so that patients can directly talk to the concerned doctors. The modules doing cancer analysis can be integrated to find how close the person associated with cancer is. This will make this web application predictable in a true sense.

7. 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 Naive 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 May 1 2020): https://www.kaggle.com/ronitf/heart-disease-uci.