REMOTE DIAGNOSIS BASED ON SYMPTOMS

Deepa Verma[*1], Ms. Kirti Kushwah[*2,] Muskan Jain[*3], Pooja Jain[*4], Riya Pal[*5]
Assistant Professor, Computer Science Engineering, Inderprastha Engineering College, Uttar Pradesh, India
Student, Computer Science Engineering, Inderprastha Engineering College, Uttar Pradesh, India
Student, Computer Science Engineering, Inderprastha Engineering College, Uttar Pradesh, India
Student, Computer Science Engineering, Inderprastha Engineering College, Uttar Pradesh, India
Student, Computer Science Engineering, Inderprastha Engineering College, Uttar Pradesh, India

ABSTRACT
Machine Learning Approach for distinctive Disease Prediction victimisation Machine Learning is based on prediction modelling that predicts illness of the patients per the symptoms provided by the users as an input to the system. This paper provides a thought of predicting multiple diseases victimisation Machine Learning algorithms. Here we are going to use the idea of supervised Machine Learning during which implementation are done by applying Decision Tree, Random Forest, Naïve Bayes and KNN algorithms which can facilitate in early prediction of diseases accurately and higher patients care. The results ensured that the system would be useful and user oriented for patients for timely diagnoses of diseases in a patient. Medicine and health care are a number of the foremost crucial elements of the economy and human life. There's an incredible quantity of change within the world we tend to live in currently and also the world that existed many weeks back. Everything has turned gruesome and divergent. During this state of affairs, wherever everything has turned virtual, the doctors and nurses are putting up most efforts to save lots of people's lives even though they need to danger their own. There are still some remote villages that lack medical facilities.
Machines are forever considered better than humans as, with none human error, they will perform tasks more expeditiously and with an even level of accuracy. A disease predictor is known as virtual doctor, which might predict the sickness of any patient with none human error. Also, in conditions like COVID-19 and EBOLA, a disease predictor is a blessing because it will determine a human's sickness with none physical contact

Keywords*: Machine Learning, Disease Prediction, Decision Tree, Random Forest, Naïve Bayes.*

I. INRODUCTION
Machine Learning is principally cope with the study of algorithms that improve with the utilization of information and experience. Machine Learning has two phases one is Training and another is Testing. Machine Learning provides an efficient platform in medical field to resolve varied aid problems at a way quicker rate. There are two forms of Machine Learning – Supervised Learning and Unsupervised Learning. In supervised learning we have a tendency to frame a model with the help of information that's well labelled. On the other hand, unsupervised learning model learn from unlabelled information.
The primary goal was to develop varied models to outline that one in every of them provides the foremost correct predictions. Many models were initiated by using varied machine learning (ML) algorithms that collected information and so divided it in accordance with gender, age group, and symptoms. The data-set was then processed in many Machine Learning models like Decision trees, Naive Bayes, KNN and Random Forest. While processing the data, the input parameters data-set was provided to each model, and also the disease was received as an output with dissimilar accuracy levels.

The intent is to deduce a satisfactory Machine Learning algorithmic program that is efficient and correct for the prediction of disease. During this paper, the supervised Machine Learning concept is employed for predicting the diseases. The main feature will be Machine Learning in which we'll be using algorithms such as Decision

Tree, Random Forest, Naïve Bayes and KNN which is able to facilitate in early prediction of diseases accurately and higher patient care.

II. LITERATURE REVIEW

There is numerous work that has been done related to disease prediction system using different Machine Learning algorithms and achieved different results for different methods in medical field.

The paper [1] "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review", summarize some of the current research on predicting heart diseases using data mining techniques, analyses the various combinations of mining algorithms used and conclude which techniques are effective and efficient. The Decision tree algorithm provides the highest accuracy 97% and hence used for prediction of the diseases.

The paper [2] "Comparing different supervised machine learning algorithms for disease prediction," identify the key trends among different types of supervised machine learning algorithms, and their performance and usage for disease risk prediction. These algorithms have a wide range of applications, including automated text categorisation, network intrusion detection, junk e-mail filtering, detection of credit card fraud, customer purchase behaviour detection, optimising manufacturing process and disease modelling .We found that the Support Vector Machine (SVM) algorithm is applied most frequently followed by the Naïve Bayes algorithm. From the studies where it was applied, Random Forest showed the highest accuracy, i.e., 53%. This was followed by SVM which is 41% of the studies that was considered.

The paper [3]"Identification and Prediction of Chronic Diseases Using Machine Learning Approach", proposed a method of identification and prediction of the presence of chronic disease in an individual using the machine learning algorithms such as convolutional neural network and K-Nearest Neighbour. Their proposed system collects structured and unstructured data obtained from various sources. Then the training data set is trained with the machine learning algorithms such as CNN and KNN to a number of times for improving the accuracy of the prediction results. For evaluating the proposed disease prediction system, four performance evaluation metrics are used. The confusion matrix consists of the true positives (TP),the true negatives (TN),false positives (FP), and false negatives (FN).This paper proposed the variations in the prediction accuracies of the four algorithms such as the Naïve Bayes, decision tree, logistic regression, and the proposed CNN and KNN algorithms as 52%, 62%, 86%, and 96%, respectively. This shows that the proposed system achieves the highest accuracy of 96% when compared to the other machine learning algorithms

The paper [4] "Heart Disease Prediction using Machine Learning Techniques", summarize that the Cardio-Vascular diseases are the primary cause of death worldwide over the past decade. Out of these deaths 80% is attributed to coronary artery disease and many more. One of the major challenges faced in the world of medical sciences today is the provision of quality service and efficient and accurate prediction. Further problem can be solved by automation with the help of Data Mining and Machine Learning techniques. The basic target of this research is to analyse the performance of various classification algorithms and in doing so find the most accurate algorithm for predicting whether a patient would develop heart disease or not. The overall focus is to define the various data mining techniques useful in effective heart disease prediction. The most Efficient and accurate prediction with a minimum number of attributes and tests is the aim of this research. The overall data were pre-processed and then used in the model for efficiency. The Random Forest with 86.88% and XG Boost with 78.68% are the most efficient algorithms. However, K-Nearest-Neighbour (KNN)performed with the worst accuracy with 57.84%.

The paper [5] "I.H. Machine Learning Algorithms, Real-World Applications and Research Directions", conducted a comprehensive overview of machine learning -based solutions, it opens up a promising and admirable direction and it also can be used as a research guide for potential research and applications for both academia and industry professionals as well as for decision-makers, from a technical point of view.

The paper [6]"Prediction of Diseases in Smart Health Care System using Machine Learning", focused on how the data mining techniques are used along with the machine learning to predict the diseases based on the user symptoms. This paper analysed the most significant risk factors of Heart Diseases of patients by extracting

multimodal features and predicting the occurrence of heart diseases using different classification techniques comparatively.

In the paper [7] "Disease Prediction System using data mining techniques" discussed about the data mining techniques like association rule mining, classification, clustering to analyse the different kinds of heart based problems.

The paper [8] "Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively" used Machine learning in project for disease prediction system. A web/android application is deployed for user for straightforward movableness, configuring and having the ability to access remotely wherever doctors cannot reach.

The paper [9] "Requirement of Machine Learning Predictive Models in the Chronic Disease", evaluates the studies associated with the diagnosis of chronic diseases in which deep learning can play a crucial role in the interpretation of chronic diseases. AI techniques like ML, cognitive computing and deep learning may play a critical role in the interpretation of chronic diseases and many more.RF and NN models are found to have the highest accuracy.

The paper [10] "Disease Prediction Using Machine Learning Over Big Data", propose a CNN-MDRP algorithm for a disease prediction from a large volume of hospital's structured and unstructured data. Using a machine learning algorithm (Naive- Bayes) Existing algorithm CNN-UDRP only uses a structured data but in CNN-MDRP focus on both structured and unstructured data the accuracy of disease prediction is more and fast as compared to the convolutional neural network and UDRP. By combining the structured and unstructured data the accuracy rate can be reach to 94.81

The paper [11] "Intelligent Prediction Methods and Techniques Using Disease Diagnosis in Medical Database: A Review" discussed different prediction algorithms of data mining used in the field of medical diagnosis. Heart disease, Diabetics, Blood disease, cancer and infertility data processing using different data mining algorithms are compiled and compared. Each machine learning and data mining algorithm algorithms performs its respective role effectively. The comparison results support the development of different hybrid data mining technologies for the more accuracy in processing of clinical data.

The paper [12] "Survey on Data Mining Algorithms in Disease Prediction", discussed that the data mining is the process of extracting hidden interesting patterns from massive database. The Medical domain contains the heterogeneous data in the form of texts or characters, numbers and images that can be mined properly to provide variety of useful information for the physicians. These patterns obtained from the standard medical data can be useful for the physicians to detect diseases, predict the survivability of the patients after disease, severity of diseases etc. This paper analysed the application of data mining in medical domain and some of the techniques used in disease prediction.

The paper "[13]" Remote Diagnosis using Machine Learning", analyses the symptoms provided by the user as input and gives the probability of the disease as an output Disease Prediction is done by implementing the Decision tree Classifier.There are various decision tree algorithms like ID3, C4.5, C5.0 and CART.Though Naïve Bayes saw significant increase in accuracy due to discretization, Random Forest gave the highest accuracy for our dataset.

III. METHODOLOGY

The project is based on multiple disease predictions in accordance with symptoms entered by patient. The initial task is to determine the problem statement, and after that making the dataset ready to work on. After that we gestate our data using scatter plot, distribution graph, etc. it is a method to find out anomalies, missing values, etc. on our data and make our dataset perfect for prognosis. And at the end, the main feature will be Machine Learning in which we will be using algorithms such as Decision Tree, Random Forest, Naive Bayes and KNN which will predict the disease for better prediction and better patient care. For this model, we have used python language as a platform to execute our Machine Learning algorithms. We have also designed an elegant GUI to provide the user the best experience.

3.1. Decision Tree

Decision Tree is one of the most widely used and practical and effective methodology for supervised learning algorithm. Decision trees are constructed via an algorithmic approach that identifies ways to separate a data set that is based on different scenarios. Decision Tree is a non-parametric and supervised learning algorithm that is widely used for both classification and regression problems. The focus of this model is to form a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The decision rules are normally in form of if-then-else. The deeper the tree is the more complex the rules and fitter the model will be. A decision tree is a tree-like graph with nodes representing the place where we pick an attribute and ask a question, edges represent the answers of that question, and the leaves constitute the actual output or class label. They are used in non-linear decision making with basic linear decision surface. Decision trees clusters the examples by sorting them down the tree from the root to leaf node, with the leaf node providing the classification to the example. Each and every node in the tree works as a test case for some attribute, and each edge descending from that node corresponds to one of the possible answers to the test case. This process is recursive in nature and is replicated for every sub tree rooted at every new node.

3.2. Random Forest

Random Forest is a popular algorithm in machine learning that belongs to the supervised learning technique. It can also be used for both Classification and Regression problems. It is based on the concept of ensemble learning, which is a process of amalgamating multiple classifiers to solve a complex problem and to upgrade the performance of the model. Random Forest is a classifier that sways a number of decision trees on different subsets of the described dataset and takes the average to enhance the imminent accuracy of that dataset. Possibly relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The eminent number of trees in the forest leads to higher accuracy and fend of the problem of over fitting. It works greatly and shows an excellent execution over other tree-based algorithms Random Forest produces good results over actual problems mainly due to massive noise in the dataset and is not based on over fitting.

3.3. Naïve Bayes

It is a machine learning algorithm for classification problems and is based on Bayes' probability theorem. The primary use of this is to do text classification which involves high dimensional training data sets. We used the Bayes theorem that can be defined as:

$$P(h|d) = \frac{P(d|h).P(h)}{P(d)}$$

Where P(h | d) is the probability of hypothesis (h) given the data (d). This is called the posterior probability. P(d | h) is the probability of data (d) given that the hypothesis (h) was true. The probability of hypothesis h being true (regardless of the data) is P(h). This is called the probability of (h). The probability of the data (regardless of the hypothesis) is P(d).

3.4. KNN

K Nearest Neighbour algorithm falls under the category of Supervised Learning algorithm and is used for classification and regression. It is an algorithm also used for imputing missing values and resampling datasets. It considers K Nearest Neighbour to predict the class or continuous value for the new data points. It is a method that implies it does not consider any assumption on underlying data. KNN is also a lazy learner method because model not learned using training data prior and the learning process is postponed to a time when prediction is requested on the new instance. It is also an instance-based learning algorithm where we do not learn weights from training data to predict output (as in model-based algorithms) but use entire training instances to predict output for unseen data. KNN is used to find the distance between new data point and each training point using distance function.
In order to find the nearest neighbors, we will calculate the Euclidean distance. The Euclidean distance between two points with coordinates (x, y) and (a, b) in the plane is given by:

$$dist(d) = \sqrt{(x - a)^2 + (y - b)^2}$$

IV. DATASET AND MODEL DESCRIPTION

In this section we are going to intricate the dataset which is used to train the Machine Learning model in this project. The dataset we have used for this project is in the structured format. The dataset which is being used consist of all the names of diseases with its respective symptoms. Since the system is based on supervised Machine Learning algorithm, the dataset is labelled with 0 or 1. After this, we have divided the dataset into two phases i.e., training dataset and testing dataset. We have trained our models using training dataset and then we applied our all Machine-Learning algorithms to this training dataset to get trained Machine Learning model. At last, we have provided the testing dataset to this trained model to test the accuracy of model.

V. SYSTEM ARCHITECTURE

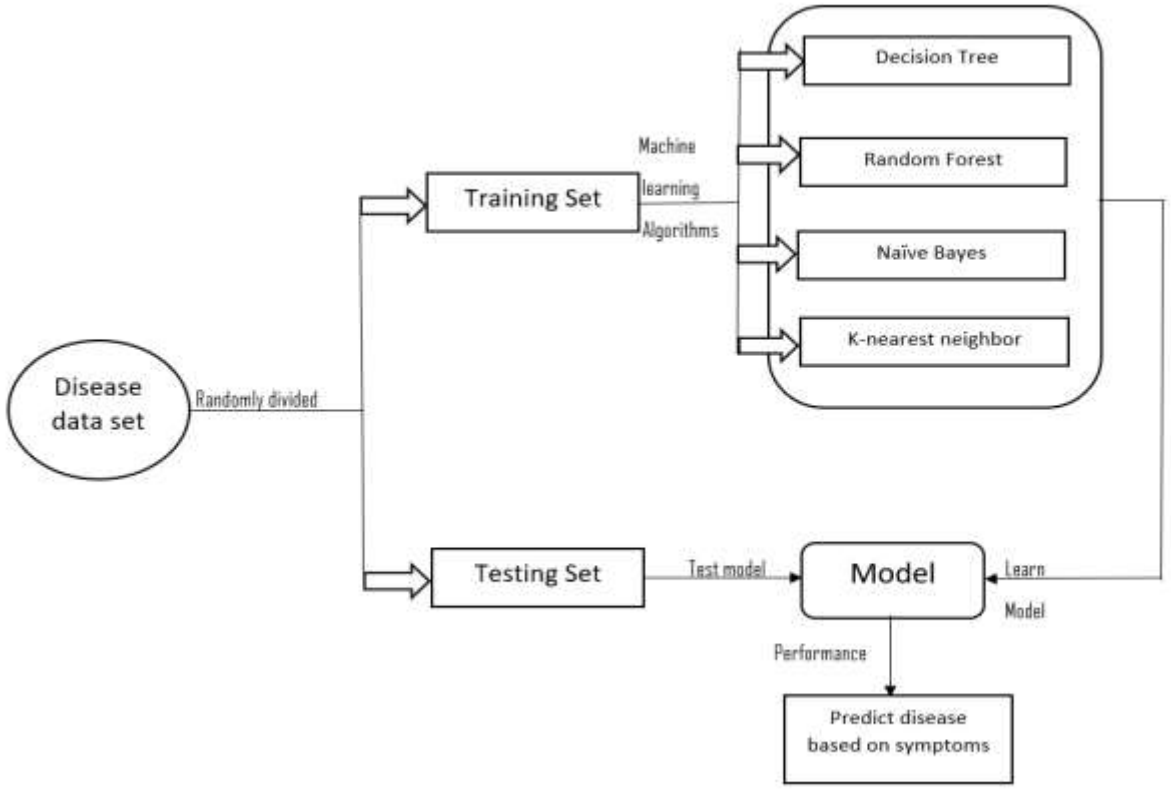

Figure 1: Architecture of proposed system

VI. RESULT

After all the above discussions, the proposed system results are obtained by implementing various Machine Learning algorithms. The Machine Learning classification techniques are namely decision Tree, Random Forest, Naïve Bayes and KNN are implemented using Python programming. The highest accuracy is delivered by Random Forest. The Random Forest classifier effectively anticipated the result with an accuracy of 95.69% approx.

For evaluating the proposed system, four performance evaluation metrics are used. The confusion matrix consists of the true positives (TP), the true negatives (TN), false positives (FP), and false negatives (FN). The confusion matrix can be graphically represented as:

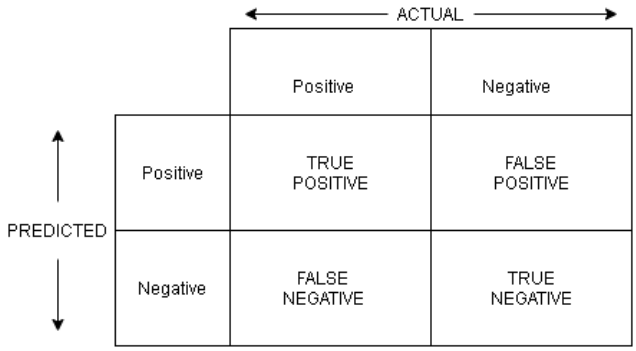

Figure 2: Confusion Matrix

Classification report includes the values for accuracy, precision, recall and F1-score. These are explained as:

6.1. Accuracy

The classification accuracy is described as the ratio of correct predicted values to the total predicted values and is depicted mathematically as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

6.2. Precision

The precision or positive predictive value (PPV) is the number of correctly identified positive results divided by the number of all positive results and is depicted mathematically as follows:

$$PRECISION = \frac{\text{TRUE POSITIVES (TP)}}{\text{TRUE POSITIVES (TP)} + \text{FALSE POSITIVES (FP)}}$$

6.3. Recall

The recall or sensitivity or true positive rate (TPR) is the number of correctly identified positive results divided by the number of all samples that should have been identified as positive and is depicted mathematically as follows:

$$RECALL = \frac{\text{TRUE POSITIVES (TP)}}{\text{TRUE POSITIVES (TP)} + \text{FALSE NEGATIVES (FN)}}$$

6.4. F1-Score

The F1 score (also F-score or F-measure) is a measure of a test's accuracy. It is calculated from the precision and recall of the test.F1 score can also be described as the harmonic mean or weighted average of precision and recall positive and is depicted mathematically as follows:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precison + Recall}$$

The accuracy is the important parameter since the prediction result is the important factor for the patient, and if it is wrong, then it will be a detriment to them. The other parameters such as precision, recall, and F1-score are for the evaluation of the model performance as shown in table below:

Table 1: Performance evaluation comparison.

	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	94.12	93	95	94
Random Forest	95.69	91	96	93
KNN	90.41	94	92	93
Naïve Bayes	91.29	90	93	91

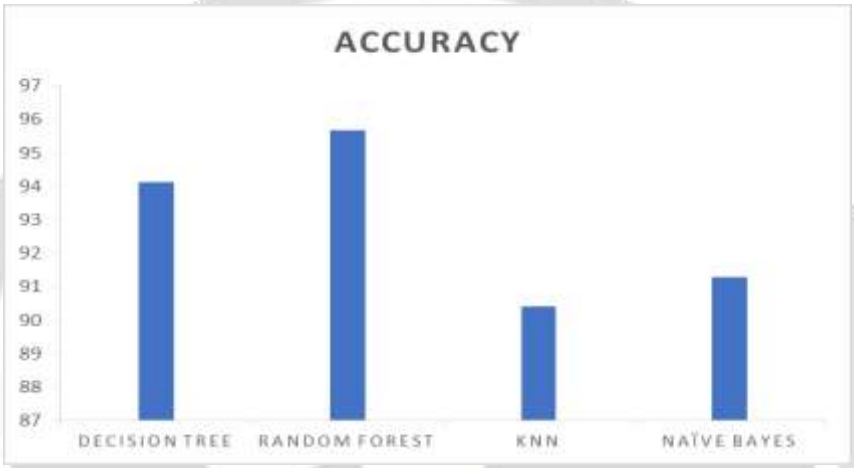

Figure 3: Comparison of accuracies of proposed algorithms.

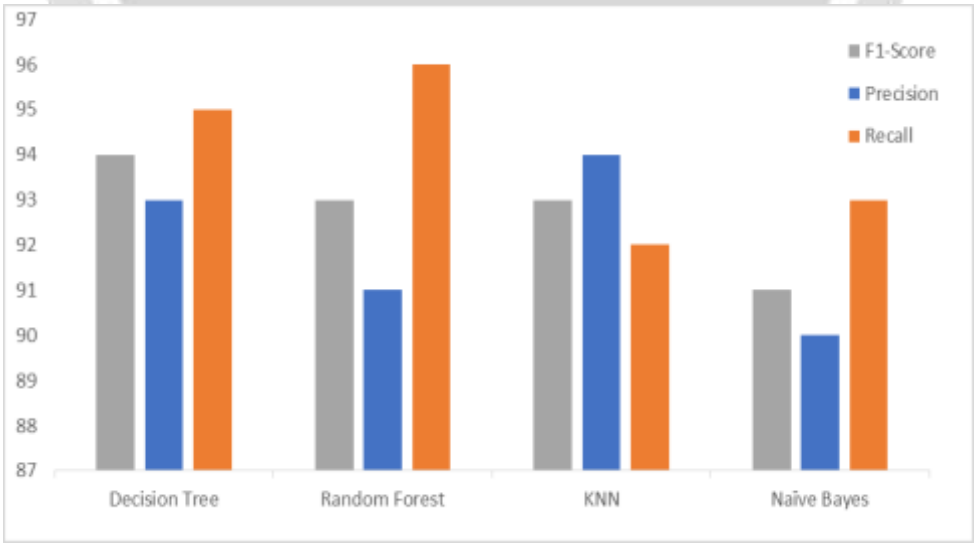

Figure 4: Comparison of other performance evaluation metrics of proposed algorithms.

VII. CONCLUSIONS

The principle aim of this paper is to predict the disease in accordance with symptoms put down by the patients with proper implementation of Machine Learning algorithm. In this paper we have used four Machine Learning algorithm for prediction and achieved the mean accuracy of more than 95% which shows remarkable rectification and high accuracy than previous work and also makes this system more reliable to understand than the existing one for this job and hence provides better satisfaction to the user in comparison with the other one. It also reserves the data entered by the user and the name of the disease the patient is suffering from in the Database which can be used as past record and will help in future also, for future treatment and thus contributing in easier health management. We have also created an elegant GUI for better interaction with the system by users which is very easy to operate. This paper shows that the Machine Learning algorithm can also be used to predict the disease easily with different parameters and models. In the end we can confidently say that our system has no threshold of the users because everyone can use this system.

VIII. REFERENCES

[1] Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee, "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review", Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10pp. 2137-2159, 2017

[2] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1–16, 2019.

[3] Rayan Alanazi, "Identification and Prediction of Chronic Diseases Using Machine Learning Approach", Journal of Healthcare Engineering, vol. 2022, Article ID 2826127, 9 pages, 2022.

[4] Pooja Anbuselvan, "Heart Disease Prediction using Machine Learning Techniques",INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY , VOLUME 09, NOVEMBER 2020.

[5]Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160 (2021).

[6]N. Shabaz Ali, G. Divya Prediction of Diseases in Smart Health Care System using Machine Learning,International Journal of Recent Technology and Engineering , January 2020

[7]Sarthak Khurana, Atishay Jain, Shikhar Kataria, Kunal Bhasin, Sunny Arora, "Disease Prediction System", International Research Journal of Engineering and Technology.

[8]Godse, Rudra A., Gunjal,Smita S., JagtapKaran A .,Mahamuni ,Neha S., &Wankhade, Prof. Suchita. (2019). Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively.International Journal of Advance Research in Computer and Communication Engineering, 8(12), 50-52.

[9]Battineni, Gopi., Sagaro,Getu.Gamo. ,Chinatalapudi, Nalini. , &Amenta,Francesco. (2020),"Application Of Machine Learning Predictive Models in the Chronic Disease", International of Personalised Medicine, 10(21), 1-11.

[10]Prof. Shubhangi Patil, Shraddha Subhash Shirsath, (2018),International Journal of Innovative Research in Science, Engineering and Technology, "Disease Prediction Using Machine Learning Over Big Data", Vol. 7, Iss. 6, pp 6752-6757.

[11]M. Durairaj, and Nandhakumar Ramasamy, 2016, Intelligent Prediction Methods and Techniques Using Disease Diagnosis in Medical Database: A Review, I J C T A, 8(5), 2015.

[12]V.Kirubha, S.Manju Priya "Survey on Data Mining Algorithms in Disease Prediction". International Journal of Computer Trends and Technology (IJCTT) V38 (3):124-128, August 2016.

[13]Raj H. Chauhan, Daksh N. Naik, Rinal A. Halpati, Sagarkumar J. Patel4, Mr. A.D.Prajapati, "Disease Prediction using Machine Learning", International Research Journal of Engineering and Technology (IRJET) (2020).

[14]Mounita Ghosh, Md. Mohsin Sarker Raihan, M. Raihan, Laboni Akter, Anupam Kumar Bairagi, Sultan Alshamrani,Mehedi Masud, A Comparative Analysis of Machine Learning Algorithms to Predict Liver Disease, Intelligent Automation & Soft Computing.

[15] Parvathi I, Siddharth Rautaray, ―Survey on Data Mining Techniques for the Diagnosis of Diseases in Medical Domain, International Journal of Computer Science and Information Technologies, Vol. 5 (1), 838-846, ISSN: 0975- 9646, 2014.

[16]Aiyesha Sadiya, Differential Diagnosis of Tuberculosis and Pneumonia using Machine Learning (2019).[17] S. Patel and H. Patel, "Survey of data mining techniques used in healthcare domain," Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53-60, March, 2016.

[18] Md. Ehtisham Farooqui, Dr. Jameel Ahmad,Disease Prediction System using Support Vector Machine and Multilinear Regression", International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN: 2347-5552, Volume- 8, Issue- 4, July- 2020

[19] Anitha,Dr.S.,& Sridevi,Dr.N. (2019). Heart Disease Prediction Using Data Mining Techniques.*Journal of analysis and Computation, 13*(2), 48-55.

[20] Bindhika,Galla Siva Sai., Meghana,Munaga., ReddyManchuriSathvika. , &Rajalakshmi. (2020). Heart Disease Prediction Using Machine Learning Techniques. *International Research Journal of Engineering and Technology, 7*(4), 5272-5276.

[21] Pingale,Kedar., Surwase, Sushant., Kulkarni,Vaibhav.,Sarage ,Saurabh., &Karve, Prof. Abhijeet .(2019). Disease Prediction using Machine Learning.*International Research Journal of Engineering and Technology, 6*(12), 2810-2813.

[22] Chauhan Raj H., NaikDaksh N. ,Halpati,Rinal A., Patel,Sagarkumar J. , &PrajapatiMr. A.D. (2020). Disease Prediction using Machine Learning.*International Research Journal of Engineering and Technology, 7*(5), 2000-2002.