An analysis of ML-based multiple disease prediction system

Abhishek Gawade[1], Prajkata Chowk[2]
[1]*Keraleeya Samajam's Model College, Khambalpada Road, Thakurli, Dombivli (East), Kanchangaon, Maharashtra, India*
[2]*Guide, Keraleeya Samajam's Model College, Khambalpada Road, Thakurli, Dombivli (East), Kanchangaon, Maharashtra, India*

Abstract-In recent years, people face various diseases because of environmental changes and their lifestyles. As a result, predicting diseases at an earlier stage becomes an important responsible task. However, reliable diagnosis based on symptoms was challenging for doctors, and the most difficult challenge is to accurately predict the disease. To overcome such problem data mining plays a significant role to predict the disease. There are many existing machine learning models available for healthcare analysis that will be focusing on a single disease at a time, like one for diabetes analysis, and one for heart disease like that. There is no single standard system where one analysis can perform more than one disease prediction based on symptoms and some other parameters like insulin level, blood pressure, etc. In this proposed project, a single standard disease prediction system is proposed which will analyze multiple diseases at a time. This project aims to detection of trends and the prevention of disease transmission, by using predictive analytics in healthcare which will enhance healthcare quality and reduces the burden on doctors. This research paper was carried out to analyze the relevant attributes, factors, and most efficient algorithms among different algorithms used in disease prediction.

Keywords: multiple disease prediction, disease prediction, healthcare, data mining, machine learning.

1.INTRODUCTION

In the past few years, the whole world faces multiple diseases because of the extremely changing environment and changes in our lifestyle. It is difficult to detect diseases at earlier stages resulting in a significant increase in mortality. The responsible task was to detect such diseases at an earlier stage which saves many lives. Doctor face problems to detect diseases earlier and more accurately on the bases of symptoms. Machine learning is a significant study of algorithms and statistical model that computer uses to perform a task without using explicit instructions, relying on pattern and inference instead. Machine learning is also used in healthcare to advance in their technique that can provide better service to the patient. The disease prediction system predicts diseases accurately based on symptoms and help doctors to precise prediction. Prediction and analysis of the diseases can facilitate the government to control the diseases and maintain a healthy and stable environment. This project can build multiple disease prediction systems that consist of multiple disease prediction models you can check any of the diseases among them by adding the specified parameter for a particular disease.

1.1. What is a disease prediction system?

A disease prediction system will predict diseases based on symptoms and some other parameters, which help doctors recognize patient health to improve the medical treatment given to the patient. Machine learning algorithms will detect the patterns of certain diseases with patient healthcare records and doctors can develop customized treatments and prescribe medicines for that specific disease in individual patients.

1.2. Advantages of disease prediction system:

Machine learning-based disease prediction systems have proven beneficial to the healthcare industry. Here are some applications of disease prediction in the healthcare industry to better engage with the users.

1.Personalizing treatment: • It allows a healthcare organization to deliver personalized patient care by analyzing patients' medical history, symptoms, and tests. Doctors can develop customized treatments and

prescribe medicines that targeted the specified diseases in an individual patient. Healthcare organizations can have access to analysis based on electronic health records for the patients. which helps the doctors to make a faster decision on what kind of treatment suits the patient.

2.Detecting fraud in an insurance claim: • The healthcare business may use machine learning models to detect invalid claims before they are paid for and to speed up the approval, processing, and payment of genuine claims. Apart from detecting insurance fraud, it will also prevent the stealing of patient data.

3.Detecting diseases in earlier stages: • There are various diseases that you need to detect in earlier stages to identify the treatment plan and assist the patient in securing a healthy way of life. Combinations of machine learning algorithms provide better assistance to doctors in the earlier detection of diseases.

4.Analyzing errors in prescription: • If a doctor chooses the wrong drug or is confused in dosing units, in such cases ml technologies may be a lifesaver in such circumstances. It analyses historical health record data and compares new prescriptions against it. The system will increase the quality of care by preventing drug overdosing and health risks.

5.Clinical decision support system:

It is an interactive computer program designed to assist health professionals with a decision – making tasks. The clinician interacts with software utilizing the clinician's knowledge as well as the software's knowledge to make a better analysis of patient data. The system generates ideas for the physician to review, and the clinician selects helpful information while discarding incorrect suggestions.

6.Drug discovery and creation: • It can discover new drugs that offer great economical value for pharmaceuticals, hospitals, and new treatment avenues for patients. It helps the process of drug creation faster and is extremely cost-effective.

7.Automating image diagnosis: • Hospitals and clinics use ML to find abnormalities in different types of medical images, such as MRI or radiological scans. Image recognition assists doctors in the diagnosis of liver and kidney infections, and tumors, improving cancer prognosis and more.

1.2. Disadvantages of the diseases prediction system:

1.Patient's safety: • The decision made by machine learning algorithms completely relies on data that has been learned. If input data is unreliable or wrong the result will be wrong as well. The flawed decision can harm the patient or even causes their death.

2.lack of quality of data: • the output you get from the machine learning algorithm depends on the quality of data passed into them. The algorithm generates suggestions for the doctor to look over, and the clinician selects important material while removing incorrect ideas. There is gas in record maintenance, inaccuracies in profile, and other difficulties. So before applying the ML algorithm, we need to spend time on information gathering, cleaning, validating, and structuring data for its purpose.

3.Privacy concern: • Another challenge of implementing AI and ML in healthcare is that collected data contain some sensitive or confidential information. It requires additional security measures to be implemented. So, it's important to look for the right ML software development company that can offer several security options to ensure your customer data is appropriately handled.

4.Problem Statement: • As we can see there are many applications for the prediction of specific diseases but there is no single application for the prediction of multiple diseases. Multiple disease predictor applications predict the disease by selecting the symptoms given in the list, it will predict the disease based on the symptom. But since many diseases have many common symptoms, there is a possibility of misdiagnosis which has many consequences on the patient's health. So in this project, we are developing an application that contains multiple diseases Users can select any disease out of multiple diseases selecting entering the symptoms parameter by the user this application will predict the diseases based on symptoms. there are fewer chances of predicting the wrong disease.

2.RESEARCH OBJECTIVE

The point of review which offers what exactly we are doing in this research paper:

- To study existing research papers for such disease web prediction applications.
- To evaluate different algorithms used in the existing system to develop such systems.

- To study precise algorithms used in disease prediction.

3.LITERATURE REVIEW

There are several existing exploration has been conducted in this field which will help us further improvement in this project. This section will elaborate on recent studies and research on new technology. They emphasize the disease prediction system using machine learning in the field of medical diagnosis.

1.Disease prediction from various symptoms using machine learning, 27 July 2020 This research paper mainly focuses on developing such a system that can able to medical diagnoses based on machine learning algorithms. The researcher will design the disease prediction system by using multiple algorithms. The system will predict more than 230 diseases based on symptoms and some other parameters such as the age and gender of a particular person system will predict the disease that he or she is suffering. For developing disease prediction researcher will study K-NN, Gaussian and Kernel Na¨ıve Bayes, Weighted KNN, and Decision tree algorithm to develop the model. All the model gives good accuracy but out of all algorithm, the weighted KNN model gave the highest accuracy of 93.5% for prediction on the above parameter. Some models will give low accuracy for the above parameter. This system provides medical resources for the treatment after the disease will be predicted.

2.Disease prediction using machine learning, December 2020 Researchers analyze different algorithms used in different disease predictions such as heart, kidney, breast, brain, etc. researcher will find that Support Vector Machine, Random Forest, and Linear regression algorithms were mostly used in prediction. While accuracy was the most used performance metric. CNN model will be most adequate for common diseases furthermore SVM shows superiority in accuracy for kidney diseases because it will handle high dimensional data and for breast cancer, Random forest is superior in the probability of correct classification of diseases it scales large data set and avoid overfitting. finally, Linear regression is the most reliable for heart disease prediction. Researcher says that we need to create a more complex algorithm to increase the efficiency of disease prediction. They suggest dataset should be expanded on multiple dimensions to avoid overfitting and increase the accuracy of the model and relevant feature selection will enhance the performance of the model.

3.Popular deep learning algorithms for disease prediction, 3 Aug 2022 This paper focused on the review of a deep learning algorithm in disease prediction. Researchers study structured algorithms including ANN, FM-Deep Learning algorithms, and unstructured algorithms including CNN, RNN, etc. algorithms. The researcher will analyze all the algorithms used in disease prediction algorithms and some problems in the existing system. In the end, researchers provide two approaches for the development of disease prediction systems in the future. The researcher will study current development, existing problems, and future trends in disease prediction algorithms with the goal of medical development trends.

4.THE ALGORITHM USED IN DISEASES PREDICTION

4.1Decision Tree:

The decision tree algorithm belongs to the supervised learning algorithm which is used for both classification and regression. It is mostly preferred for solving a classification problem. In a decision tree, two nodes are the root node (decision node) and the leaf node. Decision nodes make any decision and have multiple branches, and leaf nodes are the output node of the decision node which is not further divided. The root node after which split into the dominant input feature and then it will again split. This process is continued till all the input placed.

Advantages of the decision tree:

- It follows the same process which is a human follows while making any decision in real life.
- It helps solve the decision-based problem.
- It will think about all the possible outcomes for a particular problem.
- It requires less data cleaning compared to other algorithms.

The disadvantage of the decision tree:

- It contains a lot of layers, which makes it complex.
- It may have an overfitting issue which is overcome by using the Random Forest algorithm.

4.2Random Forest:

The random forest algorithm belongs to the supervised learning technique which is also used for classification and regression problems in ML. it is based on ensemble learning which combing multiple classifiers and contains several decision trees. It will take the prediction from each tree and based on the majority votes of prediction it will predict the final output. A greater number of trees in the forest leads to higher accuracy to solve the problem of overfitting.

Advantages of the random forest:

- It can perform both classification and regression.
- It can handle large data sets with high dimensionality.
- It increases the accuracy of the model and prevents overfitting.

Disadvantages of the random forest:

- Although it is used for classification and regression, it is not more suitable for regression tasks.

4.3Support Vector Machine (SVM):

Support vector machine is one of the supervised learning algorithms which is used for classification as well as regression. The goal of SVM is to create a decision boundary that can segregate ndimension space into classes so that we can easily put the new data point incorrect category. The decision boundary is called a hyperplane. SVM chooses extreme points that help in creating hyperplanes these extremes are called support vectors. • Type of Support Vector Machine (SVM)

1. Liner SVM:

– it is used for linearly separable data which means if data set can be classified into two classes by using a straight line.

2. Non-Linear SVM:

– it is used for non-linearly separable data which means if a dataset cannot be classified with a straight line.

4.4 K-Nearest Neighbour (KNN):

K-Nearest Neighbour is the simplest supervised learning technique. The k-NN algorithm assumes the similarity between new cases and available cases and puts new cases into the category that is most similar to available categories. K-NN stores all data points based on similarity. When a new data point appears then it can be easily classified into well suite category.

Advantages of K-NN:

- It is easy to apply and resistant to noisy training data.
- It can be more effective if the training data is large.

Disadvantages of K-NN:

- It is always necessary to identify the value of K, which might be difficult at times.
- The computation cost is high because of calculating the distance between data points for all training samples.

4.5. Linear regression:

It is a statistical method that is used for prediction analysis. Linear regression predicts continuous or numeric variables such as sales, salary, age, product price, etc. it shows the linear relationship between the dependent and one or more independent variables hence called linear regression. it finds the value of the dependent variable changes according to the value of an independent variable.

Types of regression:

1.Simple Linear Regression:

Single independent variables s used to predict the value of a dependent variable; such an algorithm is called simple linear regression.

2.Multiple Linear Regression:

More than one independent variable is used to predict the value of the dependent variable such an algorithm is called multiple linear regression.

5.METHODOLOGY

In this research, analytical and descriptive analysis where the purpose is to see what people's opinions are and understanding of disease prediction system. Here

we are analysing what is disease prediction and what kind of algorithm will be used for the disease prediction system to conclude. collect data regarding what will people think about disease prediction and whether the system will predict disease accurately or not, which is useful for deducting load on doctors or not we use google Forms for this survey.

5.1Data collection: public survey

We are collecting data through the public survey where a variety of people would use any disease prediction system or not, there are aware of various facts about disease prediction as well.

5.2Questionnaire:

Disease prediction Google form questions:

1. Do you use disease prediction application to predict disease?
2. Does it predict the disease accurately?
3. Do you know any program that automates the process of disease prediction?
4. Do You think that a computer program could automate the disease prediction process?
5. If yes, can it predict the disease accurately on its own?
6. Do you know about such automation being implemented in the medical field?
7. How do you believe this application for automatically disease prediction would be useful in day-to-day life?
8. Do you believe it will function well with various diseases?
9. Do you think the use of disease prediction software could reduce time spent on analyzing the symptoms of disease?
10. Can the burden on doctors be lessened by utilizing a disease prediction system?
11. Do you think this disease prediction automation process can lessen the risk of a fake report being generated?

5.3Result:

1.33.9% of respondents who were asked if it would be using the disease prediction method to predict their disease agreed for it would be, while 66.1% thought that the does not use disease prediction.

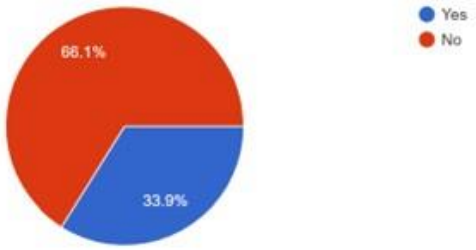

Figure 1: Caption

2.13.6% of respondents agreed that the disease prediction system will predict disease accurately, 74.6% of respondents were not sure about disease prediction will be accurate, and 11.9% disagreed.

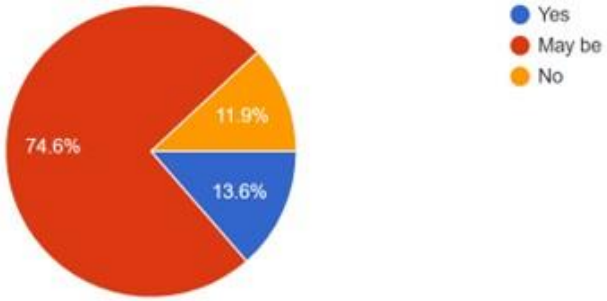

Figure 2: Caption

3.35.6 % of respondents agreed that there is some program that automates disease prediction, while 64.4% will not agree.

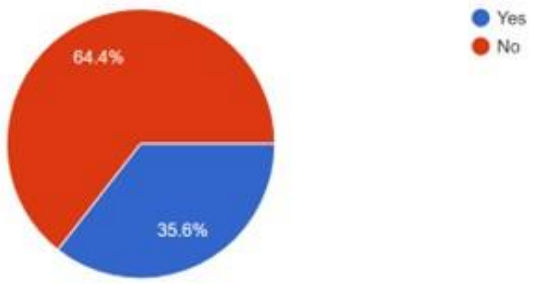

Figure 3: Caption

4.47.5% of respondents who were asked if the disease prediction will be automated using ML agreed that it would be, 44.1% of the respondent may be agreed about this may be automated, while 8.5 % will not agree.

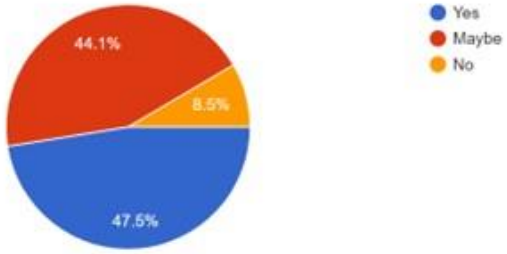

Figure 4: Caption

5.28.8% of respondents agreed that the disease prediction system will predict disease accurately, 61% of respondents may be agreed it will predict accurately, while 10.2% will not agree.

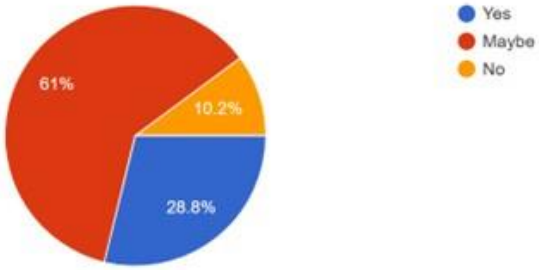

Figure 5: Caption

6.50.8% of respondents agreed that automation in the field of medicine should be implemented, while 49.2% will not agree.

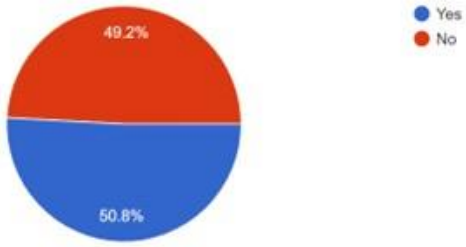

Figure 6: Caption

7.The observed answer chart for the question of whether respondents believed disease prediction would be useful in day-to-day life is as follows:

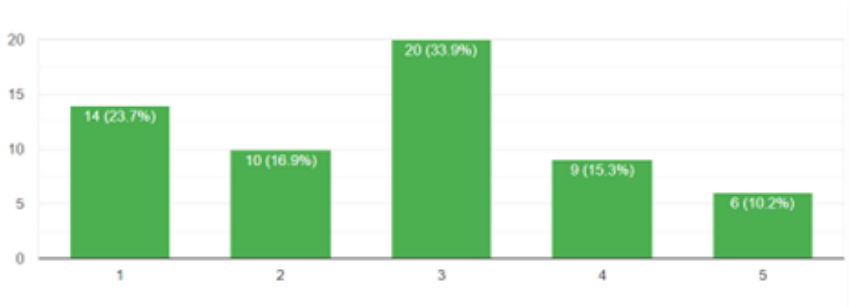

Figure 7: Caption

8.32.2% of respondents agree that it will function well for various disease predictions, 59.3% of the respondent will say that it may function well, and 8.5% of respondents will disagree.

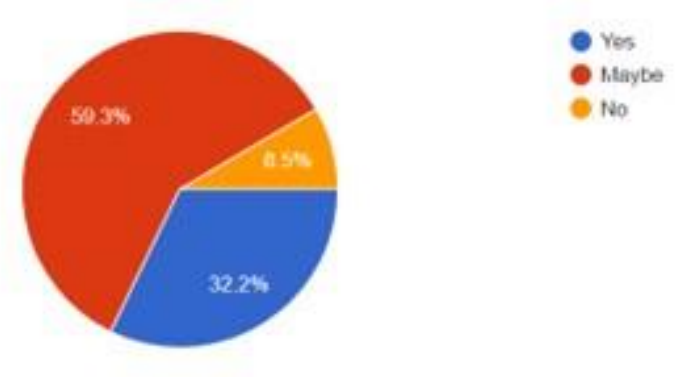

Figure 8: Caption

9.79.7% of respondents agreed that the use of disease prediction software could reduce the time spent on analyzing the symptoms of disease, while 20.3% of respondents are not agreed.

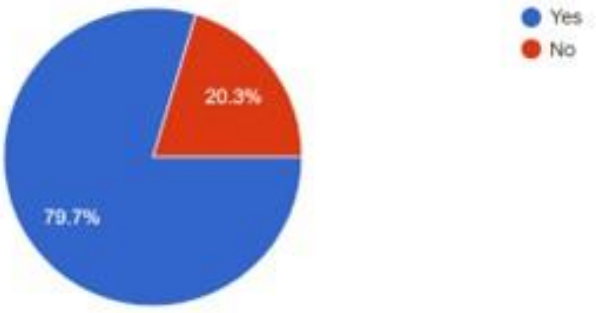

Figure 9: Caption

10.57.6% of respondents agreed that the burden on doctors be lessened by utilizing a disease prediction system, while 35.6% of respondents say it may lessen the burden on the doctors, while 6.8% of the respondents disagree.

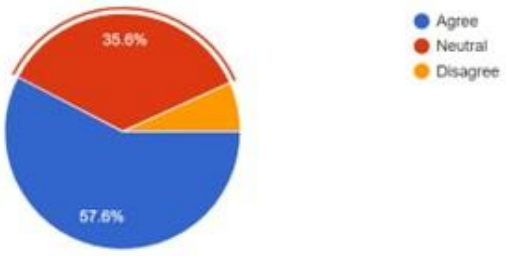

Figure 10: Caption

11.54.2% of respondents agree that the disease prediction automation process can lessen the risk of a fake report being generated, 35.6% of the respondent say it may lessen the risk, while 10.2% of respondents disagreed.

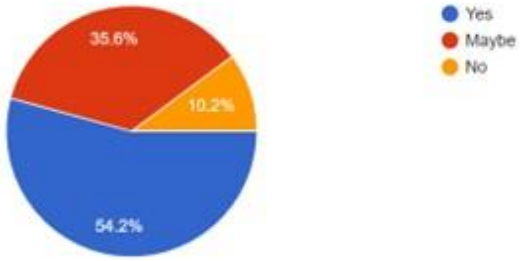

Figure 11: Caption

6.HYPOTHESIS TESTING

Hypothesis testing is a way of statistical reasoning that includes analyzing the data from the samples to drive statistical inferences to conclude population parameters or probability distribution. First, the hypothesis or assumption is a claim regarding the population parameter or probability distribution, which is known as the null hypothesis. Its id was donated by H0. After that alternate hypothesis is defined. It is donated by Ha. the alternate hypothesis is defined, as the opposite of the null hypothesis. By using sample data, the hypothesis testing technique which determines whether or not H0 may be rejected. If H0 is rejected, the statistical conclusion is that the alternate hypothesis Ha is true.

For this paper,

Null hypothesis (H0): disease prediction using a machine learning algorithm is not the best way for disease diagnosis.

Alternate hypothesis (Ha): disease prediction using a machine learning algorithm is the best way for disease diagnosis.

6.1.TEST Statistics

There are three types of tests available to determine the given assumption the null hypothesis is rejected or accepted.

The type of test is as follows:

- Chi-squared test
- T-student test
- Fisher's Z-test

For this paper, we are using two-tailed T-student tests. A t-test is an inferential statistic that determines if there is a significant difference in the means of two groups that are related in some manner.

1.Level of significance:

The chance of rejecting the null hypothesis when it is true is the significance level (also known as alpha). A significance level is 0.05 for the example, which means there is a 5% of probability of discovering a difference when there is not one. Lower significance levels indicate that more evidence is required to reject the null hypothesis.

2.Level of confidence

The confidence level indicates the probability that the location of the statistical parameter (such as the arithmetic mean) measured in the sample survey is also true for the entire population.

Level of significance = 0.05 i.e.,5%

Level of confidence = 95%

A t-score (t-value) is the number of standard deviations away from the t-mean.

Sr.No.	Data
1	33.9
2	13.6
3	35.6
4	47.5
5	28.8
6	50.8
7	23.7
8	32.2
9	79.7
10	57.6
11	54.2
Mean (X)	41.6
Standard Deviation (S)	18.54109

Table 1: Testing Data

The formula to find a t-score is:

$$t = (X - \mu)/(S\sqrt{n})$$

Where X: is the sample mean, μ: is the hypothesized mean, S: sample standard deviation, n: sample total population.

The p-value, also known as the probability value, indicates how probable your data is to have happened under the null hypothesis. Once we know of t, we can find the corresponding p-value. If the p-value is less than some alpha level (common choices are 0.01, 0.05, 0.10) then we can reject the null hypothesis and conclude that microchip is not implanted in human that is harmful to health as well as they can be hacked, tracked, and monitored by accessing the data stored on the microchip.

6.2.Calculation of T-value:

Step 1: Determine the null hypothesis and alternate hypothesis.

Null hypothesis (H0): Disease prediction using a machine learning algorithm is not the best way to disease diagnosis.

Alternate hypothesis (Ha): Disease prediction using a machine learning algorithm is the best way for disease diagnosis.

Step 2: find the test statistic.

In this case, the hypothesis mean value is 0.

$$|t| = (X - \mu)/(S/\sqrt{n})$$

$$|t| = (41.6 - 0)/(18.54109/\sqrt{11})$$

$$|t| = 7.4413959092$$

6.3. Calculating p-value:

Step 3: Calculate the test statistic's p-value. The t-Distribution table with n-1 degree of freedom is used to calculate the p-value. In this paper, the sample size is n=11, so n-1 = 11-1=10.

Level of significance =0.05

Using an online calculator will provide a p-value when the observed value is entered into the calculator. In this case, the p-value will be 0.000022.

we can reject the null hypothesis(H0) at the significance level of 0.05 because your p-value is less than 0.05. Therefore, we have enough information to conclude that Disease prediction using a machine learning algorithm is the best way to disease diagnosis.

7.CONCLUSION

In this research paper, we analyze current technologies and the different algorithms used in disease prediction. We proposed a multiple disease prediction system based on a machine learning algorithm. Here we used the Random Forest algorithm and CNN algorithm for disease prediction based on symptoms. Random forest classifier algorithm used for classification and regression problems in machine learning. RF is a classifier that contains multiple decision trees it will take the prediction from each tree on the bases of the majority votes of prediction, and it will predict the final output. If it works on classification then it will take the majority of predictions from the decision tree, and when it works on regression it will take the average of predictions from the decision tree. It is an extension of bagging, and a concept based on ensemble learning which combines multiple classifiers to solve a complex problem and improve the performance of the model. Greater the number of trees in the RF algorithm, the higher its accuracy and problem-solving ability. CNN algorithm used for image-based disease prediction.

REFERENCE

[1] Disease prediction from various symptoms using machine learning, 27 July 2020 Rinkal Keniya, Aman Khakharia, Vruddhi Shah, Vrushabh Gada, Ruchi Manjalkar, Tirth Thaker, Mahesh Warag, Ninad Mehendale. K. J. Somiya

[2] Disease prediction using machine learning, December 2020 Marouane Fethi Ferjani, Computing Department Bournemouth University, Bournemouth, England.

[3] Popular deep learning algorithms for disease prediction, 3 Aug 2022 Zengchen Yu, Ke Wang, Zhibo Wan, Shuxaun Xie, Zhihan Lv.

[4] https://appinventiv.com/blog/machine-learning-in-healthcare/