e-ISSN: 2582-5058
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:11/April-2023 Impact Factor- 6.752 www.irjmets.com

Disease Prediction using Machine Learning

Aman Patel[*1], Anurag Yadav[*2], Bhomik Gahlot[*3],
Chandransh Singh Chouhan[*4]
[*1,2,3,4]Acropolis Institute Of Technology And Research, Department Of Computer Science, Indore, India.

ABSTRACT
There is a growing importance of healthcare and pandemic has proved that healthcare is an important aspect of an individual life. Most of the medical diagnoses require going to the doctor and fixing appointments for a consultation and sometimes to get accurate disease indications we have to wait for blood reports also we have to travel long distances to seek doctor consultation. When we are not feeling well the first thing we do is to check our temperature to get an estimate or baseline idea of our fever so we can consult our doctor if the temperature is high enough similarly a medical disease prediction application can be used to get a baseline idea of disease and can indicate us whether we should take immediate doctor consultation or not, or at least start some home-remedies for the same to find temporary relief. Combining machine learning with an application interface to interact with users provides opportunities for easy interaction with the users with the machine learning model to get more accurate predictions. Sometimes people feel reluctant to visit a hospital or consult a doctor for minor symptoms but there are cases where these minor symptoms may be indications of severe health problems hence medical disease prediction maybe useful to get a baseline prediction or estimation of disease in such cases.
Keywords: Student Details, Verification, Gathering Information.

INTRODUCTION
Humans are now afflicted with a variety of ailments as a result of the current state of the environment and their lifestyle choices. It is critical to detect and forecast such diseases at an early stage to prevent them from progressing to their final stages. Most of the time, doctors find it challenging to precisely identify ailments by hand. There are multiple techniques in machine learning that do predictive analytics on large amounts of data that are used in a variety of industries. Predictive analytics assist doctors in making accurate decisions regarding health and treatment of a patient based on the vast data available, but altogether it is a very difficult process. Several diseases like cancer, diabetes etc. are becoming the cause of deaths globally and most common reason is the lack of early detection of these diseases. The lack of good medical infrastructure and a low ratio of doctors to the population is also a major factor. Based on the recommendation of WHO, the ratio of doctors to patients should be 1:1000 but in India the ratio is 1:1456, which indicates the shortage of doctors in India. Thus to save a lot of lives, early recognition and diagnosis of these diseases is very crucial. This work is all about predicting diseases using machine learning algorithms. With the advancement of technology, the better computing power and availability of datasets on open-source repositories have further increased the access to the data for use of machine learning. As a result, machine learning is being used in healthcare abundantly. Large amounts of data such as images, patient data are produced in the healthcare sector which helps to identify patterns and make predictions. Machine learning is used in healthcare to solve various real world problems that requires time-constraints and expertise of an individual.

The purpose of this study is to identify and predict patients suffering from more prevalent ailments. This might be accomplished by employing cutting-edge machine learning techniques to ensure that the categorization accurately identifies those who have diseases. Disease prediction is a difficult endeavor as well since there are variations of disease for the same type of symptoms. The proposed system uses a machine-learning algorithms known as Multinomial Naive Bayes, Random Forest Classifier, K-Nearest Neighbors to predict the disease in the data-set and provide a broad disease prognosis based on the patient's symptoms.

PROBLEM STATEMENT
The traditional diagnosis approach entails a patient visiting a doctor, undergoing many medical tests, and then reaching a consensus. This process is very time-consuming. This project proposes an automated disease prediction system to save time required for the initial process of disease prediction that relies on user input. The user gives input to the system and system provides the user with a set of probable diseases.

e-ISSN: 2582-5058
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:11/April-2023 Impact Factor- 6.752 www.irjmets.com

METHODOLOGY

A. Data collection and Data analysis

1. Dataset: The data is collected from Columbia University and this study of diseases and their corresponding symptoms is available on Kaggle.
2. Training Data: Training data is also known as training datasets, training sets, and training sets. It is an important aspect of the machine learning model which helps us to make accurate predictions and perform the tasks we want. Simply put, training data forms a machine learning model and tells you what the awaited result looks like. The model iteratively analyzes the dataset to understand its attributes precisely and make appropriate changes to enhance the performance.
3. Testing Data: The test dataset is a subset of the training dataset used to make an objective evaluation of the final model.
4. *Balanced Data*

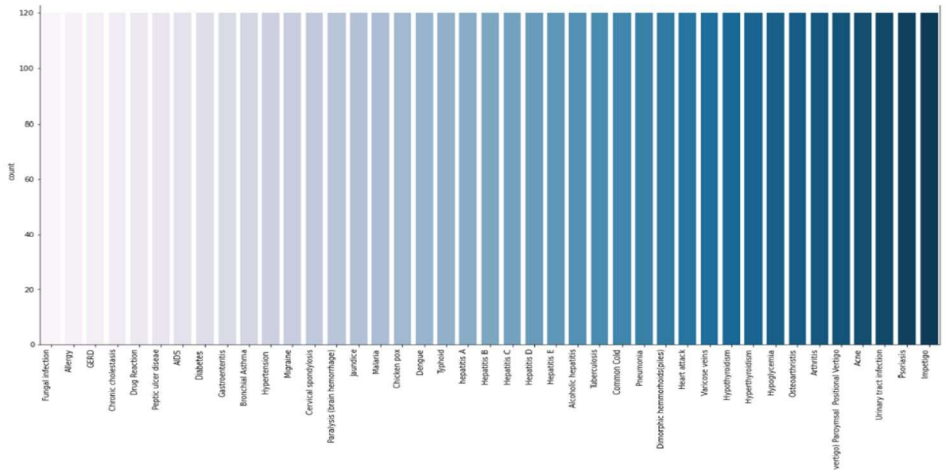
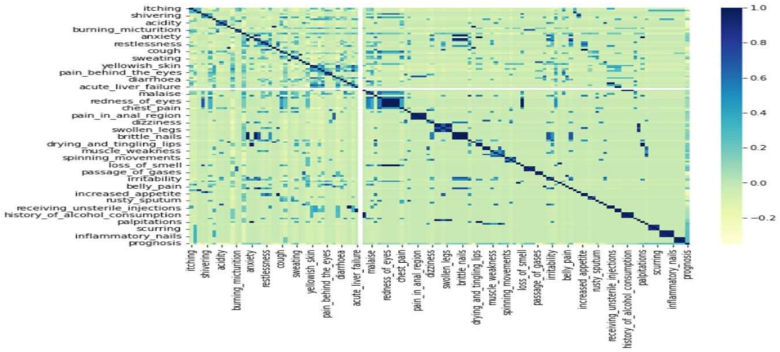


Fig. 1 Balanced data

In supervised machine learning, it is important to train an estimator on balanced data ,so the model is evenly informed on all classes. The observation of the dataset and its visualization leads us to the conclusion that the data is balanced and there's no imbalance in the data, which means that training and testing will give real exactness.

a. *Correlation of Disease with Respect to their Corresponding Symptoms:* Matrix data structures are used when there are multiple variables and the aim is to find the correlations between all these variables and store them using the applicable data structure. Thus this matrix is known as a correlation matrix. A correlation matrix is a table that shows the correlation portions between a set of variables. The pair of variables which are closely related are determined using correlation matrix. It can also be used to identify relationships between variables that may not be immediately apparent. Thus Correlation matrices are a tool for researchers and analysts who want to



understand the relationships between multiple variables.

5. Fig. 2 Correlation Heatmap.

6.

e-ISSN: 2582-5058
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:11/April-2023 Impact Factor- 6.752 www.irjmets.com

B. Algorithms and their significance

1. *Multinomial Naive Bayes:* The naive Bayes classification strategy works with Bayes' theorem, which is known to be used in arithmetic and computer science probabilistic analysis. This algorithm considers each property as an autonomous property that contributes to the final classification. For this reason, it has been praised in various studies for its accuracy and unwavering quality.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Multinomial Naïve Bayes Expression

It is a straightforward method for building classifiers, which are models that give class labels to problem cases represented as vectors of characteristic values. Here the class labels are selected from a finite set. For training such classifiers, there is no one algorithm, but rather a variety of algorithms based on the same principle: all naive Bayes classifiers assume that the value of one feature is independent of the value of any other feature, given the class variable.

For example, if the fruit is red, round, and around 10 cm in diameter, it is termed an apple. A naive Bayes classifier examines each of these characteristics to contribute independently to the likelihood that this fruit is an apple, regardless of any possible confounding variables.

2. *Random Forest Classifier:* The decision tree is the basic building block of random forest classifiers. A decision tree is a hierarchical structure created from a data set's characteristics (or independent variables). The decision tree is divided into nodes based on a measure connected with a subset of the characteristics. The random forest approach was used to create the prediction model, and the results were compared to those of a decision tree based on multiple logistic regression and a classification and regression tree. The recognition rate was used to calculate the model's forecast accuracy. Random forests are multi-decision tree ensemble classifiers that train multiple decision trees at random. The random forest approach is made up of two steps: a training step that creates numerous decision trees and a test step that classifies or predicts an outcome variable based on an input vector. Forest F = f1,..., fn represents the ensemble form of random forest training data .After averaging the distributions obtained from each forest's decision trees by T (the number of decision trees), classification was performed. The mean for continuous target variables and the majority vote for categorical target variables were used to aggregate the predictors of each sample.
3. *K-Nearest Neighbors Classifier:* The KNN method is a supervised machine learning technique which can be used to handle classification and regression issues. K-Nearest Neighbors works by calculating the distances between a query and all of the instances in the data. The K closest examples to the query is picked and the most frequent label is voted. The nearest neighbor is the data point that is the smallest distance in the feature space from the new data point. Also, K is the number of such data points to consider in the implementation of the algorithm. Therefore, when using the ANN algorithm, the distance metric and the K value are two important considerations. All data points in the training dataset are considered to predict bin / continuous values for new data points. Finds the "K" nearest neighbor (data point) of a new data point from the feature space and its class label or continuous value. The class labels assigned to most of the K-nearest neighbors from the training dataset are considered to be the prediction classes for the new data points.
4. *Support Vector Machine (SVM):* SVM (Support Vector Machine) is a supervised machine learning technique which can be used to solve classification and regression problems. It is, however, mostly employed to solve classification /classification difficulties. In n-dimensional space each data item is plotted, with the value of each feature being the value of a certain coordinate in the SVM algorithm. Then we accomplish classification by locating the hyper-plane that clearly distinguishes the two classes or more than two classes. SVM is a model-free method for solving classification problems that does not make any assumptions about the distribution or interdependency of the data. The SVM technique has the potential to outperform traditional statistical methods like logistic regression in epidemiologic

e-ISSN: 2582-5058
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:11/April-2023 Impact Factor- 6.752 www.irjmets.com

studies and population health surveys, especially when multivariate risk factors with small effects (e.g., genome-wide association data and gene expression profiles), limited sample size, and a lack of understanding of underlying biological relationships among risk factors are present. This is especially true in the case of prevalent complex diseases, where several risk variables, such as gene-gene interactions and gene-environment interactions, must be incorporated in order for prediction models to have significant discriminative strength.

7. Parameters used for Calculating the Accuracies of the Machine Learning Models

The confusion matrix is an N x N matrix used to evaluate the performance of the classification model. Where N is the number of target classes. The matrix compares the actual target value with the value predicted by the machine learning model. This gives you a complete picture of the classification model's performance and error types.

a. Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

For a good classifier, the recall should ideally be 1 (high). If the numerator and denominator are the same, that is, the recall is 1. H. TP = TP + FN, which also means that FN is zero. As FN increases, the denominator value becomes larger than the numerator and the recall value decreases

b. Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

The accuracy of a good classifier should ideally be 1 (high). The accuracy is 1 only if the numerator and denominator are the same, that is, TP = TP + FP. This also means that the FP is zero. As FP increases, the denominator value becomes larger than the numerator and the precision value decreases.

The F1 score will only be 1 if both the fit and recall are 1. The F1 score is high only if both the fit and recall are high. The F1 score is the harmonic mean of precision and recall and is a better measure than precision.

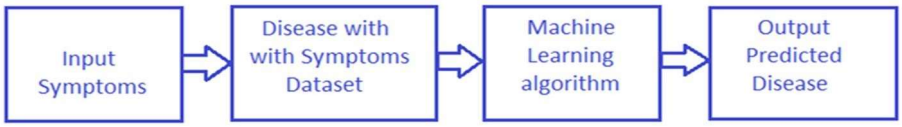


Fig. 3 Block Diagram of flow of the system

i. Comparison of all algorithms

1. Multinomial NAÏVE BAYES

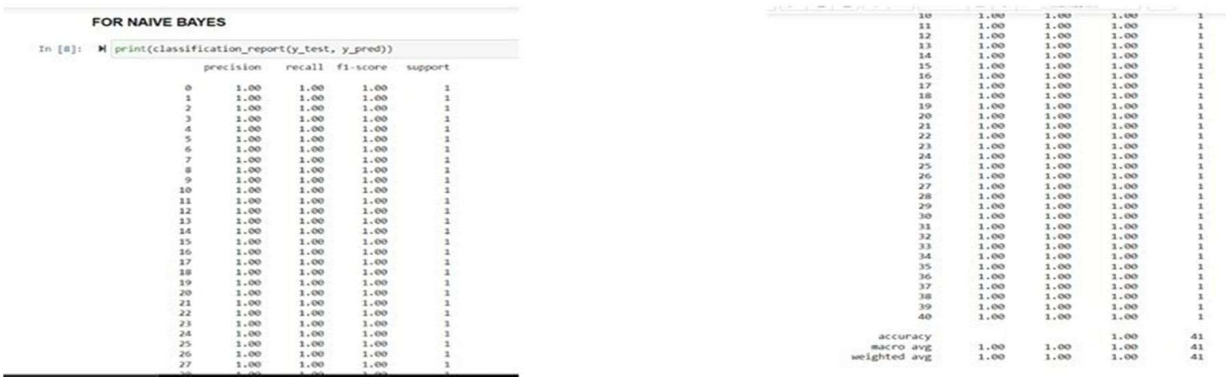


Fig. 4 Classification report Naïve bayes

e-ISSN: 2582-5058
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:11/April-2023 Impact Factor- 6.752 www.irjmets.com

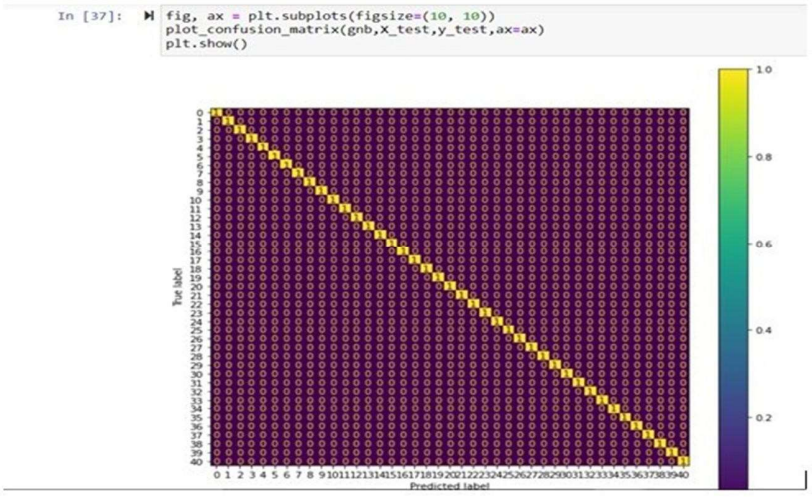
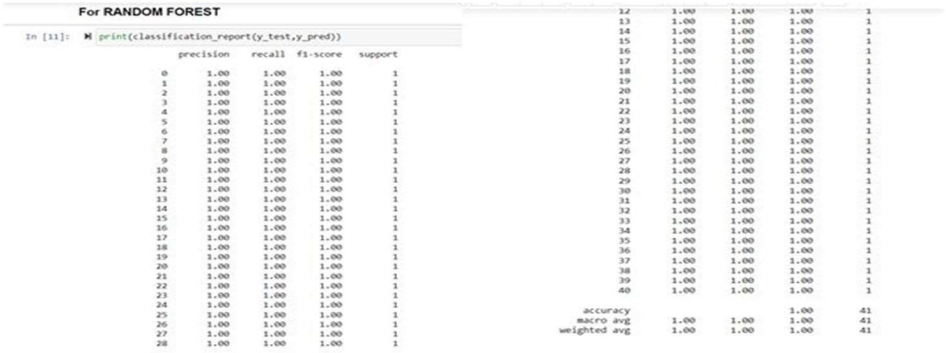


Fig. 5 confusion matrix naïve bayes

The naïve bayes model fits the training data and gives an accuracy of 100% on all parameters in the classification problem.

2. Random Forest Classifier



8. Fig. 6 Classification report for random forest

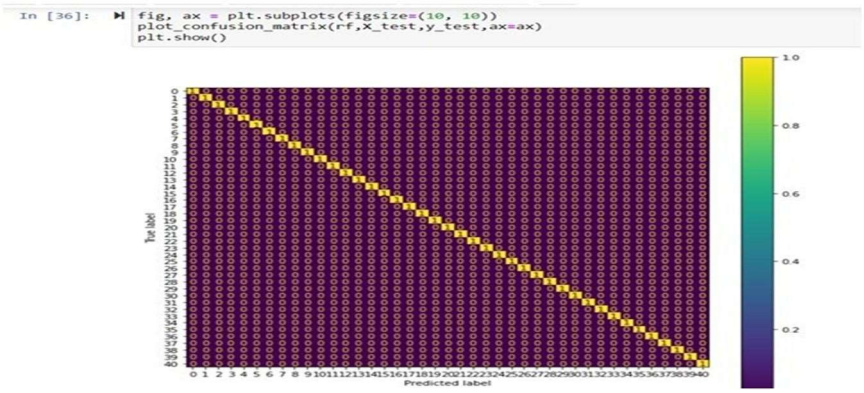


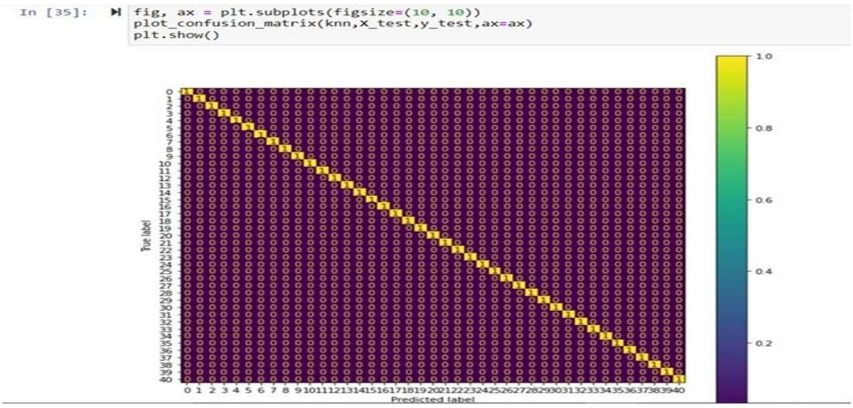
Fig. 7 Confusion matrix for random forest

The random forest classifier model fits the training data and gives an accuracy of 100% on all parameters in the classification problem.

e-ISSN: 2582-5058
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:11/April-2023 Impact Factor- 6.752 www.irjmets.com

3) K-Nearest Neighbor Classifier

Fig. 8 Classification report on KNN



The k-nearest neighbors model fits the training data and gives an accuracy of 100% on all parameters in the classification problem.

The naïve bayes model fits the training data and gives an accuracy of 100% on all parameters in the classification problem.

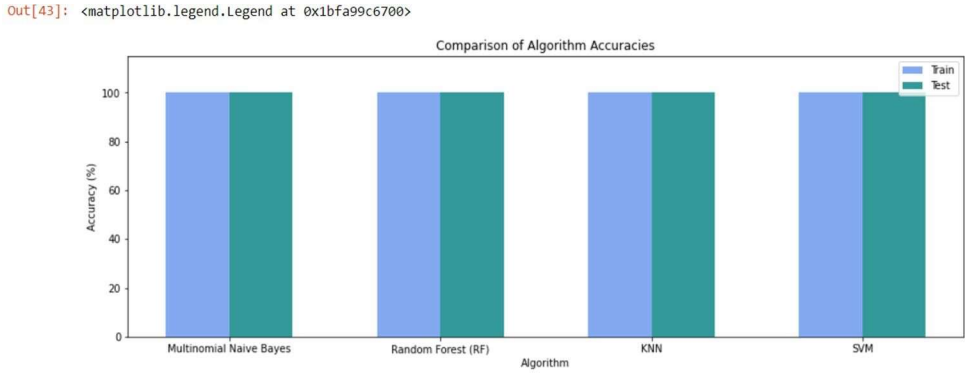


Fig. 12 Accuracies of algorithms

e-ISSN: 2582-5058
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:11/April-2023 Impact Factor- 6.752 www.irjmets.com

B. Factors that May Affect Accuracies.

1. *Selection of less symptoms:* The project is set up in such a way that the user is introduced to the chatbot system and a preference is asked whether the user would like to get an estimation or prediction of his/her disease based on the input into the chatbot the user is taken to the prediction system that contains machine learning algorithms like Naïve Bayes, Random Forest Classifier ,K- Nearest Neighbors and Support Vector Machine classifier which takes the user's symptoms as input and the prediction labels based on individual accuracies. The user has the option of selecting one to five symptoms, if less symptoms are entered, the accuracy will be lower hence, the larger the number of symptoms, the more accurate the diagnosis.
2. *Selecting of symptoms which are not relevant to each other:* If the user selects the symptoms which are not relevant to each other in the diagnosis of the disease then the machine learning models may give less accuracy on the classification problem.

CONCLUSION

This paper proposed a method of identification and prediction of the presence of a disease in an individual using the machine learning algorithms like Naïve Bayes, Random Forest Classifier, K-Nearest Neighbors and Support Vector Machines given that the user have given the maximum of five symptoms. Thus, the given methods also lead to a comparative analysis of various machine learning algorithms for multiclass classification. It is highly believed that the proposed system can reduce the risk of diseases by diagnosing them earlier and also reduces the cost of diagnosis, treatment, and doctor consultation, however the selection of symptoms does play an important role in the accuracy of the disease prediction

REFERENCES

M. A. Jabbar and S. Samreen, "Heart disease prediction system based on hidden naïve bayes classifier," 2016 International Conference on Circuits, Controls, Communications and Computing (I4C), 2016, pp. 1-5, doi: 10.1109/CIMCA.2016.8053261.

Keniya, Rinkal and Khakharia, Aman and Shah, Vruddhi and Gada, Vrushabh and Manjalkar, Ruchi and Thaker, Tirth and Warang, Mahesh and Mehendale, Ninad and Mehendale, Ninad, Disease Prediction From Various Symptoms Using Machine Learning (July 27, 2020). Available at SSRN: https://ssrn.com/abstract=3661426 or http://dx.doi.org/10.2139/ssrn.3661426S.

S. Grampurohit and C. Sagarnal, "Disease Prediction using Machine Learning Algorithms," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-7, doi: 10.1109/INCET49848.2020.9154130.R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.

Kaur, Simarjeet, et al. "Medical diagnostic systems using artificial intelligence (ai) algorithms: Principles and perspectives." IEEE Access 8 (2020): 228049- 228069.

Bhavsar, Kaustubh Arun, et al. "Medical diagnosis using machine learning: a statistical review." Computers, Materials and Continua 67.1 (2021): 107-

125. FLEXChip Signal Processor (MC68175/D), Motorola, 1996.

Chen, Min, et al. "Disease prediction by machine learning over big data from healthcare communities." Ieee Access 5 (2017): 8869-8879.

Jadhav, Saiesh, et al. "Disease prediction by machine learning from healthcare communities." International Journal of Scientific Research in Science and Technology (2019): 29-35.

Pingale, Kedar, et al. "Disease prediction using machine learning." International Research Journal of Engineering and Technology (IRJET) 6 (2019): 831-833.