# A Literature Survey on Performance Analysis of Machine Learning Algorithms for the Prediction of Diseases

Aditi Mondal *1, Anupama Ghosh 2, Samriddhi Sardar 3, Joy Adhikary4
Department of Computer Science
Bijoy Krishna Girls' College, Howrah
5/3 Mahatma Gandhi Rd., Howrah, West Bengal 711101, India
*1aditimondal561@gmail.com, ,2samriddhi.ss31@gmail.com,3anupamaghosh87655@gmail.com,
4adhikari.joy04@gmail.com

***Abstract*- Disease prediction using machine learning approaches is one of the widely used applications in the healthcare industry. Due to people's lifestyle and the environmental circumstances, health issues are increasing day by day. So, the early prediction of disease has become a very important part of our daily life. Besides that, the proper prediction of disease is a serious matter. This literature provides a survey on various machine learning algorithms in case of recognize the disease. These algorithms are trained from the symptoms of the user data. Efficacy of algorithms tested based on their accuracy in case of disease prediction ability. Several machine learning algorithms, like k-Nearest Neighbour (k-NN), Naive Bayes, Decision Tree, and Random Forest have been used on several real life dataset.**

***Keywords: Machine Learning, Classifier, Random Forest, Decision Tree.***

## I. INTRODUCTION

Machine learning is a subset of artificial intelligence (AI). It can learn from previous user-provided data, which helps to make predictions. Machine learning-based models are widely used in a variety of real-world problems. It is used in various fields such as weather forecasting, facial recognition, disease prediction, and speech recognition. In this study, we analysed the performance of machine learning-based models for disease prediction

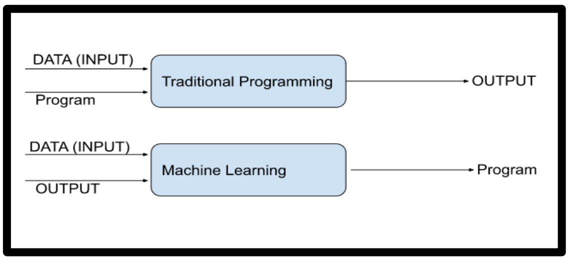

Fig. 1: Diagram of General approach of Machine Learning

Recognizing the appropriate disease from patient's symptoms and history is not a simple task. However, healthcare problems can also be solved or simplified with machine learning techniques. Efficiently we can detect the disease using full machine learning concepts [5]. In [2] People always want to learn new things, especially since internet usage is increasing every day. When a problem occurs, many people will want to look it up on the web. Hospitals and doctors have less access to the Internet than the general public. When people get diseases, they don't have many options. As a result, this device can be beneficial. Smart Health System is a project providing end-user support.

If someone has already contracted the disease, they will need to see a doctor, which is time-consuming and expensive. It can also be difficult for the user if the doctor or hospital cannot be contacted due to an unrecognised illness. Therefore, if the method described above can be performed with automated software that saves time and money, the process may be easier for the patient.

Intelligent Healthcare System is a web-based program that predicts a user's illness based on symptoms. Datasets from various health-related websites were compiled for a smart health system. Consumers can determine possible illness based on the symptoms provided by online consulting. This document proposes a framework for enabling users to receive online health advice from intelligent health systems. Various symptoms and diseases were entered into the system. Users can share their symptoms and problems with the system. Next, analyse the user's Symptom to see if there is a disease associated with it.

This article uses intelligent data mining techniques to infer the most reliable suspected diseases that may be associated with a patient's symptoms, and an algorithm (naive Bayesian) to reduce symptoms to probabilities. Not only does this method simplify the doctor's job, it also benefits the patient by providing the care they need as soon as possible.

Currently, heart disease is mostly responsible for the number of deaths. Men are affected more in heart disease than women. Smoking and drinking habits are mostly responsible for that. Human life is primarily dependent on the function of the heart, because the heart supplies blood to all the organs of the body. Heart disease includes high blood pressure, heart attack, heart disease, and heart failure. For heart disease, it is necessary to predict disease at an early stage and start treatment early. Machine learning models bring an opportunity to detect heart disease. The study uses machine learning algorithms to predict heart disease early based on factors such as age, gender, and other algorithms such as blood pressure.

This paper is organised as follows. Section 2 describes the literature survey. Section 3 describes several algorithms of machine learning which have been used for disease prediction.

## **II.** LITERATURE SURVEY

Researchers can predict particular chronic diseases in specified regions and communities. For synt hetic data (synthetic data is a data format that mimics realworld patterns generated by machine lea rning algorithms. Various sources spot synthetic data for different grounds) This system uses vario usmachine learning algorithms like, k Nearest Neighbours, Decision Tree, Naive Bayes.

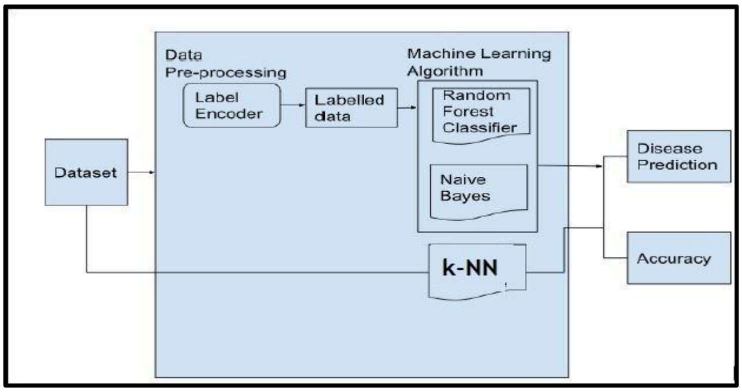

Fig. 2: Diagram of General approach of Machine Learning [4, 5]

. Figure 2 shows a system based on various machine learning algorithms, which can be used to predict the most of the chronic diseases. The system uses structured data as inputs. This system is used by end users, patients, or other users. First, patients input all the syndromes they suffer from, and then these syndromes are input into the machine learning model. Then, depending on the syndrome, the best algorithm is applied to predict disease. The system displays the status based on the machine learning technology. A simple Bayesian algorithm used to predict the disease based on the symptoms of patients. A k-NN algorithm is used for classification. Logistic regression helps extract features with the most influential values. Decision trees algorithms used to divide large input sets into smaller pieces. The outcome of the system will be the disease predicted by the model.

In [2][3], intelligent health systems are based on naive Bayes classifiers, which are used to classify diseases based on user feedback. In case of disease prediction, a web interface framework has been used. Machine Learning based system to accurately predict common diseases. The symptom dataset was imported from the UCI repository, which contains symptoms of many common diseases. The system incorporates the k-NN classifier to predict multiple diseases.

## **III.** METHODS

Several machine learning algorithms are available. Some of the algorithms described in the context of disease prediction.

### k- Nearest Neighbour (k-NN):

k-Nearest Neighbor(k-NN) [7] is a widely used machine learning algorithm. It is a simple, straightforward and versatile classification technique. Users predict illnesses in the healthcare system. The proposed method classifies disorders into several categories and uses symptoms to indicate which disease manifests. This is the best option for managing multiple jobs associated with a category. By specifying the nearest neighbor class, the k-NN classification algorithm predicts target specifications for new instances. The user must enter a value for 'k' and the best option depends on the information available [7].

The k-NN [1, 7] algorithm is a very simple technique and relatively easy to implement. This is a very useful method for solving classification problems. First, we need to choose a value for k (neighbors), then calculate the distance between the train (normal and disease sample) and test data points. After that, select those k neighbours which have minimum distance (in case of minimization problem). Finally, predict the class of test data point based on the selected k neighbour, whether it belongs to normal or disease sample class [7].

### Naive Bayes

Naive Bayes [2] is a simple supervised learning model but highly effective predictive modelling rule. The term "naive" refers to the independence conjecture, allowing the joint probabilities to be decomposed into products of marginal probabilities. A naive Bayesian classifier might detect the presence of all other features independently of other items in the class. It is very much suitable for large data sets. Bayes' theorem provides a way to compute $P(b|a)$ from$P(b)$, $P(a)$, and $P(a|b)$. Consider the formula:

$$P(b\ v\ a) = P(a\ v\ b)P(b)/P(a)$$

## Decision Tree

A decision tree [8] is a non-parametric algorithm which can be used easily on huge real datasets without involving various parametric structures. It creates a framework that can be used to elegantly partition a large collection of datasets into smaller sets [3]. It can be helpful in disease prediction systems, because it classifies symptoms to reduce the complexity of the data set. Researchers also can use decision trees [3, 8] to develop prediction models for a target instance. They trained the model with the help of useful information which is hidden in the dataset. It is used to predict the disease based on the symptoms [8].

In the caseof a large dataset, the decision tree [3, 8] algorithm works really well in considering accuracy. It is primarily used to solve classification problems in data mining. Basically, it is a tree structured classifier, where features of the real life datasets are described by internal nodes, decision rules are defined by branches and internal nodes show the outcome. Hence we have constructed the decision tree classifier model which is trained using the dataset in a shorter period by normalizing our data using standardization techniques known as case gradient descent.

## Random Forest

The random forest [6] algorithm is one of the widely used machine learning algorithms. It is a supervised learning model that contains a number of decision trees on various subsets of the given dataset. The performance of this algorithm does not depend only on a single decision tree, because it takes the prediction from each tree and accurately predicts the final outcome. Various classification and regression problems are solved with the help of it. It combined various classifiers to solve real life complex problems. Substantial growth of its popularity due to several reasons like, the very short training time compared to other algorithms, high accuracy, overcoming the problem of over fitting and so on.

The random forest [6] algorithm works in two phases. The first phase is to develop a random forest by combining decision trees. The second phase is to make predictions for each decision tree which is created in the first phase. In the case of each test data point, find all the predictions from each tree. Based on all predictions, assign the class for test data points. It is one of the most efficient methods for high-dimensional data modelling. It can handle missing values and continuous values and also provides higher accuracy than other machine learning based algorithms [5].

## **IV.** CONCLUSION

The primary motivation for disease prediction is to predict disease based on symptoms. It takes symptoms as input from the user and produces disease prediction output. Usage is simple. Users can use it anywhere in the world, if the models are fitted in a web application. Models can use several machine learning algorithms such as k-NN (k Nearest Neighbor), Naive Bayes, Decision Trees, Random Forest and so on. This research provides the summary on these algorithms. These algorithms should help future disease prediction software developers and facilitate personalized patient care.

## V. REFERENCES

[1] Charbuty, B., &Abdulazeez, A. (2021). Classification based on decision tree algorithms for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20-28.

[2] Frank, E., Trigg, L., Holmes, G., & Witten, I. H. (2000). Naive Bayes for regression. Machine Learning, 41(1), 5-25.

[3] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. SN Computer Science, 1(6), 1-6.

[4] Ali, J., Khan, R., Ahmad, N., &Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, *9*(5), 272.

[5] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., &Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, *13*(4), 18-28.

[6] Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, *13*(1), 1063-1095.

[7] Li, L., Weinberg, C. R., Darden, T. A., & Pedersen, L. G. (2001). Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, 17(12), 1131-1142.

[8] Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2), 130.