# An Analysis of Disease Prediction Algorithms

**Dr Thara L[1], Abinaya B[2]**
[1]Associate Professor, Dept of MCA
[2]Dept of MCA
[1, 2] PSG College of Arts & Science, Coimbatore, India

***Abstract-*** *The aim of using machine learning algorithms for disease prediction is to immensely help to solve health-related problems by assisting physicians in predicting and diagnosing diseases in an early phase. The Disease Prediction methodology is based on predictive modelling. This predicts the patient's disease based on the symptoms they provide as input. In this method, it analyses the patient's symptoms and returns the disease's probability as an output. The accuracy of machine learning models for disease prediction depends on a number of factors, including the quality of the training data, the choice of algorithm, and the amount of computing power available. The algorithms then identify the patterns that are associated with each disease. This allows them to predict the likelihood of a patient having a particular disease, with their symptoms and other clinical data. As technology continues to develop, it is likely to play an increasingly important role in the diagnosis and treatment of diseases. With better prediction algorithms, the medical practitioners can improve their ability to respond to emergency situations and render better quality treatment and save lives. These also help in rendering a quality treatment on an affordable cost with a well-developed predictive model. This article focuses on various disease prediction algorithms at present and analyses various metrics and their pros and cons.*

***Keywords***- ML, Artificial Neural Network(ANN), k-NN Algorithm, SVM Algorithm, Bayesian Networks, Random forest tree, Logistic Regression,disease prediction algorithm, Prediction reliability.

## I. INTRODUCTION

Not only inventing advanced medicines and treatments are enough but also the technology and care must also rely on diagnosing the disease at an early stage. Since the accuracy and variations are different with a lot of differences and similarities it is important to diagnose the disease correctly while keeping in mind the constraints and limitations of a patient. For example, A women who is breast feeding can't be suggested to take certain tablets of steroids. Thus, due to the dynamic and complicated nature involved in diagnosing of a disease there is a greater need for accurate prediction inorder to avoid loss of life and furthermore damage. Availing timely treatment and faster prediction accompanied by accuracy might prevent fatality in emergency situations. In this decade where there is growing number of diseases, powerful machine learning algorithms can be used to identify diseases prediction.While the invention of effective medicine is important there is also a need for correct diagnosis of diseases in order to avoid serious medical implications.

While the implementation of machine learning algorithms is needed there is also a need for complex testing such as aggregation, clustering and regression, variance for finding a suitable algorithm to implement in various cases and situations. For example, for treating a cancer patient an oncologist might require deep learning and past history and tracks of a patient and requires clustering methods in order to find the similarities and track further developments. In order for researchers and physicians to use the prediction results as a guideline in considering ML method for disease prediction, the main contribution of this paper is to determine the most reliable techniques in disease prediction for various diseases in an effective way.

## II.LITERATURE REVIEW

Diseases can be diagnosed while there is a change in the critical parameters deviating from homeostasis condition. The amount of variation and similarities with the conditions in the pretrained data, the machine would be able to predict the disease efficiently. Although the clinical data available is abundant but the selection and using of appropriate machine learning model is also needed for the prediction to be done accurately and effectively[1]. An effective algorithm must be able to predict almost all the clinical stages and also must be able to predict all the possible outcomes based on the response of the patient to the treatment. The algorithms below which when performed yields the best outcomes of all algorithms and tested the results of the research. The research papers analysed were from the year 2016 to 2022. On comparison we got to know the best results were given by Support Vector Machine for similarity assessment using aggregation while the Random Forest Tree method utilises Multiple Decision Trees to get more optimised result by choosing the most repeated value as the central value(modal value). Thus, there are a lot developing systemic methods in order to get more accurate predictions. Here we present our study on various algorithms

used for disease prediction, by looking onto the efficacy of outcomes based on the parameters and various critical factors. Especially there are many diseases which show the similar symptoms in such a case the machine learning models must be trained with various complicated datasets and more exposure to a variety, varsity of datasets rather than volume of datasets [2]. By taking into consideration a lot of factors and parameters such as age, weight, height, gender that is the BMI Index and geographic location of the patient the disease and its severity may vary, thus this might help in providing more effective medication and also keep a check on the spreadof zoonotic diseases across the world.

## III. METHODOLOGY

*1. Machine Learning:*

Among the many definitions of machine learning (ML), machine learning can be defined as a collection of techniques that can be used to build a model from datasets and the patterns observed by summarizing the parametric variances and correlations in order to make predictions or take actions to improve a system. In the branch of artificial intelligence referred to as "machine learning," complicated patterns are identified and conclusions are drawn from experimental datasets. Machine-based learning models can be analytical and predict outcomes, or they can be descriptive and learn from the provided training data. The initial purpose of the development of machine learning techniques was to automate the generation of data for incorporation into knowledge structures. AI techniques are of great interest to those who create intelligent surroundings, and so-called evolutionary computing has drawn inspiration from them [3].

.

A. *Supervised Learning:*

The AI network is trained to use a mapping function to map output data using a set of inputs and targets. The presence of a "mentor" and the input output data used for the training are the two most important components of supervised learning. The final two divisions are regression and classification. In supervised learning, methods including support vector machines, random forecasting, and linear regression are employed.

*B. Unsupervised Learning:*

Unsupervised learning lacks predefined direction or a mapping function and instead trains an AI network to uncover hidden patterns, responses, and distributions using an input dataset that has not been classified or labelled that is using an unlabelled dataset since mostly the labelled datasets are heavily priced. Just two of the challenges of unsupervised learning are clustering and association. K-means and auto-encoders are a couple of examples of unsupervised learning techniques.

## IV. BACKGROUND STUDY AND RELATED RESEZRCH WORK

In this today's world humans are subjected to a variety of diseases and the amount of people susceptible to it are in a surmounting rate. Thus there is a need to analyse the type of diseases and its other parameters such as similarities and differences inorder to provide appropriate medical treatment accordingly. Here Machine Learning comes into play for the decision making based on the parameters.

Even though the data available and its symptoms are clearly articulated and documentedthere is a very less intervention of computer technology in this field for diagnosis of diseases and thus proper pretrained model should be deployed using appropriate machine learning algorithms [4]. But I also agree to the fact that certain diseases require separate algorithms for further treatment and tracking of the disease in a patient. So on an overview this paper summarises on algorithms which can be used for overall disease prediction rather than on a particular field of disease

*A. Decision Tree:*

The decision tree (DT), also known as the classification and regression tree (CART), is one of the most well-known machine learning (ML) techniques. Using this approach in supervised learning, categorization-related problems can be solved. This method seeks to predict the obtained target variable's class value. As a result, Decision Tree is more adept at choosing the best suitable hypothesis from the training dataset. A categorical type class and the numerical type class are the target classes for regression trees in classification trees [5].

*B. K-Nearest Neighbour(kNN):*

This algorithm is also non-parametric. It is one of the simplest methods for regression and classification. It can be difficult to establish the k value when dealing with huge data sets. The calculation cost is high because the similarity or regression parameter (k) between the data points for each training sample must be determined. It is resilient to a variety of training data. A large amount of training data might make it more effective.

*C. Support Vector Machine (SVM):*

A supervised machine learning approach for classification and regression tasks is the support vector machine (SVM) algorithm. It is predicated on the notion of locating a hyperplane in the data that best distinguishes the various classes.

Finding the hyperplane that optimizes the margin between the two classes is how the SVM algorithm operates. The margin is the distance existing between the nearest data points from each class and the hyperplane. Support vectors are the data points that are closest to the hyperplane.

An extensively labelled collection of data points is used to train the SVM algorithm. A collection of features and a class label should be present for each data point. The SVM method will discover the hyperplane's optimal parameters.

*D. Naïve Bayes Classifier Algorithm:*

It is based on the Bayes theorem, a mathematical method for estimating the likelihood of an event occurring in the presence of another occurrence.

Given the class label, naive Bayes classifiers assume that a data point's features are independent of one another. Although this assumption is frequently unfound, it makes Naive Bayes classifiers extremely quick and simple to train.

We require a labelled collection of data points in order to train a Naive Bayes classifier. A collection of features and a class label should be present for each data point. The probability of each characteristic occurring for each class will be taught to the classifier.

*E. Random Forest Tree (RF):*

It is built on the idea of ensemble learning, it combines the predictions of various individual learners to create a forecast that is more accurate.

A huge number of decision trees are built using the random forest tree algorithm on various subsets of the training data. A random subset of features is employed at each split for training each decision tree, which is done using a sample training data chosen at random. This lessens the over fitting issue that the decision trees used to experience [6].

The random forest technique used trained decision trees to produce predictions by average their individual predictions. The class with the highest votes is predicted in classification problems.

*F. Logistic Regression:*

A supervised machine learning approach called logistic regression is used to forecast the likelihood of a particular result. It is a linear model, but it forecasts a probability between 0 and 1 rather than a continuous value. This makes it suitable for categorization jobs like determining whether a patient has a specific illness or whether a client would churn.

Algorithm for logistic regression: The logistic regression algorithm uses the training data to fit a logistic function. The input values are condensed by the logistic function, a sigmoid function, into a range between 0 and 1.

There are several ways to train the logistic regression algorithm, but gradient descent is the most used one. An iterative process called gradient descent modifies the model's parameters in order to reduce the loss function. The degree to which the model matches the training set of data is indicated by the loss function.

The probability of a specific result for a new data point can be predicted using the logistic regression model after it has been trained. This is accomplished by figuring out the logistic function’s output for the new data point.

## V. SUMMARISATION OF BRIEF CONTRIBUTION OF RESEARCHERS IN LITERATURE

*Table 1. Summary of ML Techniques on Disease Prediction*

| Reference No | Contributors | Methodology adopted | Dataset & Pros | Discussions by Authors |
|---|---|---|---|---|
| 1. | Rinkal Keniya, Aman Khakharia, Vruddhi Shah, Vrushabh Gada, Ruchi Manjalkar, Tirth Thaker, Mahesh Warang, Ninad Mehendale | kNN-93.5% | It used more than 230 diseases for further processing. It takes into account the symptoms, age, and gender of an individual, the diagnosis system gives the output as the disease that the individual might be suffering from diseases. [7] | It is vulnerable to various calculations of parameters of concurrence. It is a memory-based approach. New algorithms will continue to emerge as the field of machine learning continues to evolve. [7] |
| 2. | Sneha, Grampurohit, Chetan Sagarnal | Decision Tree classifier, Random forest classifier, and Naïve Bayes classifier | For analysis, information from 4920 patient records with diagnoses for 41 disorders was chosen. 41 diseases made up a dependent variable and 95 out of 132 independent variables (symptoms) that were highly connected to illnesses were chosen. [8] | In the future, as the amount of data continues to grow, Random Forest is likely to be used more frequently for big data analytics and feature engineering. [8] |
| 3. | Palle Pramod Reddy, Dirisinala Madhu Babu, Hardeep Kumar and Dr.Shivi Sharma | Random forest classifier-95% | The criterion considered in this study was age, sex, smoking, being overweight, and drinking alcohol, blood sugar, heart rate, and blood pressure. The risk level for the parameters were ranging from 1 to 100. (1-8). [9] | Random Forest shows higher accuracy in prediction. [9] |
| 4. | Rayan Alanaz | Naïve Bayes, decision tree, logistic regression and the KNN and CNN algorithm as 52%, 62%, 86%, and 96%. | The data set was prepared by gathering disease symptoms, a person's lifestyle, and information on doctor visits. These factors are all taken into account in this general disease prediction where the age, weight are ignored. [10] | It is firmly believed that the suggested system can lower the risk of chronic diseases by diagnosing them sooner and that it can also lower the cost of medical consultation, diagnosis, and treatment. [10] |

| | | | | |
|---|---|---|---|---|
| 5. | Kunal Takke, Rameez Bhaijee, Avanish Singh, Mr. Abhay Patil | Naïve Bayes, Random Forest Classifier, K-Nearest Neighbours and Support Vector Machines kNN-100%,SVM-100% | Determining and predicting a disease's presence in a person using machine learning algorithms including Naive Bayes, Random Forest Classifier, K-Nearest Neighbours, and Support Vector Machines, provided the user has provided a maximum of five symptoms. [11] | The user has the choice of choosing from one to five symptoms; the more symptoms entered, the higher the accuracy; the more symptoms entered, the lower the accuracy. [11] |
| 6. | K. Gaurav, A. Kumar, P. Singh, A. Kumari, M. Kasar*, T. Suryawanshi | Naive Bayes Algorithms-94.8%, weighted KNN model-93.5%, SVM model also gives a very close value, Random forest model-97% | Multiple diseases are categorized using this method based on symptoms and various geographic locations. These locations play a role in determining the outcomes because the database presupposes that certain areas have unique symptoms. [12] | Thus with this it is possible to lessen the issues the medical business faces with the absence of medical workers and the patients' inability to pay for dictators.<br><br>This can be accomplished by automatically routing patients to specialist practitioners once the disease detector machine has given the initial prediction. [12] |
| 7. | Dr Visumathi J, Tetala Durga Venkata Rama Reddy, Velagapudi Abhinandhan, Panamganti Anil Kumar | Naive Bayesian high accuracy of 91.2%. Random Forest - 85.7% and Decision Tree - 81.3% | The patient's age, gender, symptoms, medical history, and test results for a number of illnesses are all included in the dataset. The dataset is put through complex data preparation and feature selection procedures in order to guarantee the accuracy and reliability of the system. [13] | The techniques used, which included dataset selection, pre-processing, feature selection, and the Naive Bayesian network algorithm, in addition to discussing the social relevance of this work and the potential benefits of accurate disease prediction for bettering patient outcomes. [13] |
| 8. | Indukuri Mohit, K. Santhosh Kumar, Uday Avula Kumar Reddy and Badhagouni Suresh Kumar | The logistic regression gave better accuracy for diabetic diseases and breast cancer related diseases while kNN provided closer values for heart diseases. | In this various data sets of the diseases mentioned here, along with the various attributes concerning it. [14] | We can expand this research in the future by include additional diseases that are learned by machine learning models as well as diseases that use deep learning models. [14] |

| | | | | |
|---|---|---|---|---|
| 9. | Simarjeet Kaur; Jimmy Singla; Lewis Nkenyereye; Sudan Jha; Deepak Prashar; Gyanendra Prasad Joshi | fuzzy k-nearest neighbour - 80% | In this paper they had reviewed the most recent literature for the past ten years, from January 2009 to December 2019, in this report. The study took into account the eight most popular databases, where 105 papers in total were discovered. These papers underwent a thorough study in order to categorize the most popular AI methods for medical diagnostic systems. [15] | Additionally, the functions of AI ML methods for diagnostics systems utilizing sensor-based computing frameworks will also be researched. [15] |
| 10. | C K Gomathy | Decision Tree and Linear Regression Decision Tree 84.5%, Random Forest – 98.95%, Naïve Bayes 89.4%, SVM 96.49%, KNN 71.28% | Here the Grails framework was used to create Disease Predictor effectively. [16] | This method can be developed by further adding other factors and seeing their correlation among them. [16] |
| 11. | Neha Gupta Kriti Gandhi Shafali Dhall | Random Forest, SVM, naive Bayes, KNN, Classification and Regression Trees (CART), and Logistic Regression (LR) algorithms<br><br>Logistic Regression - lowest performance (accuracy score of 80.85%) | It includes the DBMI dataset, which included 133 symptoms and 42 disease types. [17] | To create a robust model, we can combine, i.e. take the average of the predictions from all three models, ensuring that even if one model predicts incorrectly while the other two do correctly, the final result will be accurate. [17] |
| 12. | Tanmay Ture, Amol Sawant , Rohan Singh , Prof. Chetna Patil | Logistic regression, Random Forest classifier, Support Vector Machine classifier<br><br>Random Forest | This paper studies the machine learning techniques to spot people with serious illnesses like heart disease, kidney disease, and diabetes early on so that the right therapies can be administered to | We can expand the current system in the future by including other ailments. [18] |

| | | | | |
|---|---|---|---|---|
| | | -98.52% accuracy on heart disease prediction, 98.73% accuracy on kidney disease prediction and 80.55% accuracy on diabetes prediction. | them. The dataset was collected for all the three diseases and the model was trained for prediction. [18] | |

## VI. CONCLUSION

Disease prediction algorithms are a valuable tool in the medical industry especially in today's rapidly growing world of both population and diseases. Thus these are important for the medical communities and research communities for development of infrastructure. These models must be further tested for increasing the accuracy and here it was observed that more combined hybrid models give higher accuracy compared to the traditional machine learning algorithms. Thus more research into this field is more beneficial for the human community and in the future its interference into the medical prediction might become inevitable.

## REFERENCES


[1] Israel Júnior Borges do Nascimento, Author Orcid Image; Milena Soriano Marcolino, Author Orcid Image; Hebatullah Mohamed Abdulazeem Author Orcid Image; Ishanka Weerasekara, Author Orcid Image; Natasha Azzopardi-Muscat Author Orcid Image; Marcos André Gonçalves Author Orcid Image; David Novillo-Ortiz: Impact of Big Data Analytics on People's Health: Overview of Systematic Reviews and Recommendations for Future StudiesImpact of Big Data Analytics on People's Health: Overview of Systematic Reviews and Recommendations for Future Studies on 19 January, 2021.

[2] Baptiste Vasey, MMed; Stephan Ursprung, MMed; Benjamin Beddoe, BSc; et al: Association of Clinician Diagnostic Performance With Machine Learning–Based Decision Support SystemsA Systematic Review on 11 March, 2021

[3] Ahmed Fadlelmoula,Susana O. Catarino, ORCID, Graça Minas, ORCID and Vítor Carvalho: A Review of Machine Learning Methods Recently Applied to FTIR Spectroscopy Data for the Analysis of Human Blood Cells on 29 May 2023

[4] David Ben-Israel, W. Bradley Jacobs, Steve Casha, Stefan Lang, Won Hyung A. Ryu, Madeleine de Lotbiniere-Bassett, David W. Cadotte: The impact of machine learning on patient care: A systematic review on 31 December, 2019.

[5] David Lyell,corresponding author Enrico Coiera, Jessica Chen, Parina Shah, and Farah Magrabi: How machine learning is embedded to support clinician decision making: an analysis of FDA-approved medical devices on April 2021.

[6] Jessica M Schwartz, Amanda J Moy, Sarah C Rossetti, Noémie Elhadad, Kenrick D Cat: Clinician involvement in research on machine learning–based predictive clinical decision support for the hospital setting: A scoping review on July 2021.

[7] Rinkal Keniya, Aman Khakharia,Vruddhi Shah, Vrushabh Gada, Ruchi Manjalkar, Tirth Thaker, Mahesh Warang, Ninad Mehendale: Disease Prediction From Various Symptoms Using Machine Learning on 8 October 2020

[8] Sneha Grampurohit; Chetan Sagarnal: Disease Prediction using Machine Learning Algorithms on June 2020

[9] Palle Pramod Reddy, Dirisinala Madhu Babu, Hardeep Kumar and Dr.Shivi Sharma: Disease Prediction using Machine Learning on 5 May 2021

[10] Rayan Alanaz: Identification and Prediction of Chronic Diseases Using Machine Learning Approach on 25 February 2022

[11] Kunal Takke, Rameez Bhaijee, Avanish Singh, Mr. Abhay Patil: Medical Disease Prediction using Machine Learning Algorithms on 02 May 2022

[12] K. Gaurav, A. Kumar, P. Singh, A. Kumari, M. Kasar*, T. Suryawanshi: Human Disease Prediction using Machine Learning Techniques and Real-life Parameters on June 2023

[13] Dr Visumathi J, Tetala Durga Venkata Rama Reddy, Velagapudi Abhinandhan, Panamganti Anil Kumar: Multi-Disease Prediction Using Machine Learning Algorithm on 04 May 2023

[14] Indukuri Mohit, K. Santhosh Kumar, Uday Avula Kumar Reddy and Badhagouni Suresh Kumar: An Approach to detect multiple diseases using machine learning algorithm on 16 September 2021
[15] Simarjeet Kaur; Jimmy Singla; Lewis Nkenyereye; Sudan Jha; Deepak Prashar; Gyanendra Prasad Joshi: Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives on 03 December 2020
[16] C K Gomathy: THE PREDICTION OF DISEASE USING MACHINE LEARNING on December 2021
[17] Neha Gupta, kritiGandhi, Shafali Dhall: Disease prediction using machine learning on September 2022
[18] Tanmay Ture, Amol Sawant, Rohan Singh, Prof. Chetna Patil: Multiple Disease Prediction System on 08 March 2023