© APR 2023 | IRE Journals | Volume 6 Issue 10 | ISSN: 2456-8880

Review on Different Algorithms for Disease Prediction

OM KHEDKAR[1], RUTUJA LABHSHETWAR[2], JAYANT KHANDEBHARAD[3], VAISHNAVI DHAKARE[4], PROF. PARAG JAMBHULKAR[5]

[1, 2, 3, 4, 5] *Department of Computer Engineering Pune Institute of Computer Technology Pune, Maharashtra*

Abstract- ***Any health-related concern must be accurately and promptly examined to prevent and treat illness. The traditional diagnostic approach might not be sufficient in the case of a serious illness. A diagnosis that is made using machine learning (ML) can be more accurate than one made using conventional methods in the construction of a medical diagnosis system for disease prediction. Supervised machine learning (ML) algorithms have shown tremendous potential in outperforming traditional systems for illness diagnosis, aiding medical personnel in the early detection of high-risk disorders. This study discusses a number of algorithms that can be employed to identify diseases based on the patient's current symptoms. We also provide a summary of the outcomes produced by the various algorithms.***

Indexed Terms- ***Machine learning, Decision Tree Classifier, Random Forest Classifier, Naïve Bayes Classifier, Support Vector Machine, K-Nearest Neighbors, ID3.***

I. INTRODUCTION

Anyone who is currently ill need to make a time-consuming and costly healthcare appointment. Because the condition cannot be diagnosed, it can also be challenging for the person if they are far from medical facilities. Therefore, we must develop a system that enables patients to identify diseases based on the symptoms they are experiencing.

Minor symptoms might often make individuals hesitant to go to the doctor or go to the hospital, but there are situations where they may be signs of more serious health issues. In these situations, medical illness prediction may be helpful to gain a general idea of the disease. The availability of datasets on open-source repositories and improved computer power brought about by technological advancements have enhanced access to data for machine learning applications. Machine learning is therefore widely applied in healthcare.

Malaria, dengue, impetigo, diabetes, migraines, jaundice, chicken pox, and other ailments have a substantial impact on a person's health and can even be fatal if left untreated. The healthcare sector can make smart decisions by "mining" the massive database they already have, or by identifying its hidden links and patterns. Various machine learning algorithms like Decision Tree Classifier, Random Forest Classifier, Naïve Bayes Classifier, Support Vector Machine, K-Nearest Neighbors Classifier can be used to build a system which can be able to predict diseases.

India's health systems face significant obstacles in terms of quality, accessibility, affordability, and inequality. On the one hand, India has some of the top hospitals in the world, which helps to fuel the expansion of the medical tourism industry. On the other hand, there is a severe lack of competent medical experts. 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 [7]. Rural locations have exceptionally low ratios, forcing patients to travel great distances for even the most basic care.

II. RELATED WORKS

The use of algorithms like Decision Tree Classifier, Random Forest Classifier, Naïve Bayes Classifier, Support Vector Machine, K-Nearest Neighbors, ID3 for disease prediction has opened possibilities that will help in accurate and premature disease diagnosis on early stage.

III. DIFFERENT ALGORITHMS USED FOR DISEASE PREDICTION

A. Random Forest Classifier:

Random Forest, as the name implies, is a classifier that uses several decision trees on different subsets of the provided dataset and averages them to increase the dataset's predictive accuracy. Instead, then depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions. Higher accuracy and overfitting are prevented by the larger number of trees in the forest.

[4] In order to classify data, their prediction system uses Random forests, an ensemble learning method of classification, regression, and other tasks that work by building a large number of decision trees during training period and then producing the class that is the mean of the categories or mean prediction of the individual trees.

[5] Random Forest typically produces excellent results even without hyper-tuning. Overfitting is a major drawback of the decision tree method, as noted in the decision tree, it seems it is overcome in random forest.

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

[8] In the case of unstructured text files, they use the random forest algorithm to automatically select features.

Model accuracy diabetes model 98.25, breast cancer model 98.25, heart disease model 85.25, kidney disease model 99, liver disease model 78.

B. Support Vector Machine:

One of the most well-liked supervised learning algorithms, Support Vector Machine, or SVM, is used to solve Classification and Regression problems. However, it is largely employed in Machine Learning Classification issues. The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. A hyperplane is the name given to this optimal decision boundary.

[1] In this study, in case of breast cancer SVM excelled in terms of accuracy and precision. Because it works well with both semi-structured and unstructured data, the SVM classifier is best for detecting renal disorders.

[2] The SVM model proved to give 79.13% accuracy as studied by Mir et al.

[3] In this disease prediction system, Multilinear Regression (MLR) is utilized to forecast the result while Support Vector Machine (SVM) is used to predict the disease. It makes an effort to forecast a value using two or more variables. MLR is a sort of regression technique used when there are numerous independent values.

[7] In epidemiologic research and population health surveys, the SVM methodology has the potential to outperform conventional statistical methods like logistic regression, particularly when multivariate risk factors with minor effects are present.

C. Decision Tree:

The most effective and well-liked technique for categorization and prediction is the decision tree. A decision tree is a type of tree structure that resembles a flowchart, where each internal node represents a test on an attribute, each branch a test result, and each leaf node (terminal node) a class label.

[4] This system makes use of the Decision Tree technique, which may be applied to both continuous and categorical dependent variables. In this algorithmic programme, the population is frequently divided into two or more homogenised sets. This is done in support of the most important characteristics and freelance variables to create the most distinct teams possible.

[5] Overfitting resulted when all 132 symptoms from the original dataset were taken into account rather than only 95 symptoms. The tree appears to memorise the dataset provided, failing to classify the new data as a result.

D. Naive Bayes:

A group of classification algorithms built on the Bayes' Theorem are known as naive Bayes classifiers. It is a family of algorithms rather than a single method, and they are all based on the idea that every pair of features being classified is independent of the other.

[2] In this study, this model gives an accuracy of 16.8% which was quite less.

[4] The Naive Bayes method, which learns the likelihood that an object with specific qualities belongs to a given group or class, is employed in this prediction system.

[7] Given the class variable, all naive Bayes classifiers make the assumption that the value of one feature is unrelated to the value of any other feature. The naïve bayes model fits the training data and gives an accuracy of 100% on all parameters in the classification problem.

E. KNN algorithm:

The K-NN algorithm makes the assumption that the new case and the existing cases are comparable, and it places the new instance in the category that is most like the existing categories. A new data point is classified using the K-NN algorithm based on similarity after all the existing data has been stored.

[2] In this study , the accuracy of Fine KNN was 80.3%. A medium KNN had a 61.8% accuracy. The accuracy of coarse KNN was 5.3%. Accuracy for Subspace KNN was 73.2%. The highest accuracy of 99.7% using the KNN model for the prediction and classification of heart diseases.

[7] The K instances that come closest to the question are chosen, and the label with the most votes wins. The data point that is the closest to the new data point in the feature space is referred to as the nearest neighbour. The k-nearest neighbors model fits the training data and gives an accuracy of 100% on all parameters in the classification problem.

A. Results published in previous studies

TABLE I. RESULTS OF DIFFERENT ALGORITHMS USED FOR DISEASE PREDICTION

Ref	Dataset Used	Algorithms	ccuracy
[1]	Dataset from UCI repository	1.SVM 2.Random Forest 3.CNN 4.LR	1. 95.85% 2. 96.27% 3. 84.50% 4. 86.89%
[2]	Custom dataset–230 diseases, symptoms - age, gender	K-nearest neighbors	93.5%
[3]	Structured dataset created by collecting patients symptoms & diagnosis from local hospital	SVM Multilinear Regression	68% to 87%
[4]	Custom dataset	1.Random Forest Classifier 2.Decision Tree Classifier 3.Naïve bayes	-
[5]	Dataset containing 132 symptoms lead to 42 diseases based on 4920 records of patients	1.Random Forest Classifier 2.Decision Tree Classifier 3.Naïve bayes	93% 93.2% 93.6%
[7]	Kaggle dataset from Columbia University	1.Random Forest Classifier 2.K-Nearest Neighbors 3.Naïve bayes 4. SVM	%
[8]	Data collection has been done from the internet to identify the disease here the real symptoms of the disease are collected	Random Forest Classifier	Diabetes Model 98.25 Breast Cancer Model 98.25 Heart Disease Model 85.25 Kidney Disease Model 99 Liver Disease Model 78

REFERENCES

[1] Ferjani, Marouane (2020) “Disease Prediction Using Machine Learning”.

10.13140/RG.2.2.18279.47521.

[2] P. Jha, T. Biswas, U. Sagar and K. Ahuja, "Prediction with ML paradigm in Healthcare System," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 2021, pp. 1334-1342, doi: 10.1109/ICESC51422.2021.9532752.

[3] 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.

[4] Sarthak Khurana1, Atishay Jain, Shikhar Kataria, Kunal Bhasin, Sunny Arora, “Disease Prediction System”, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056.

[5] Ahelam Tikotikar, Mallikarjun Kodabagi, “A SURVEY ON TECHNIQUE FOR PREDICTION OF DISEASE IN MEDICAL DATA”, 2017 International Conference On Smart Technology for Smart Nation.

[6] Sneha Grampurohit, Chetan Sagarnal, “Disease Prediction using Machine Learning Algorithms”, 2020 International Conference for Emerging Technology (INCET) .

[7] Kunal Takke, Rameez Bhaijee, Avanish Singh, Mr. Abhay Patil, “Medical Disease Prediction using Machine Learning Algorithms”, International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 10 Issue V May 2022.

[8] Palle Pramod Reddy, Dirisinala Madhu Babu, Hardeep Kumar, Dr.Shivi Sharma, “Disease Prediction using Machine Learning”, Volume 9, Issue 5 May 2021 |