www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

# Disease Predictive Modeling Using Machine Learning and Symptom Data

[1]A. RAJESH REDDY, [2]G. VENU. [3]V.V.S VINOD KUMAR, [4]B. VIJAYA DURGA

[1, 2, 3, 4] Assoc. Professor, Krishna Chaitanya Degree College, Nellore, AP, India.

***Abstract*— In medical analysis, computer-aided diagnosis (CAD) is a rapidly developing and multifaceted field of study. Since medical diagnosis errors can lead to highly misleading medical treatments, a lot of work has been done in the last several years to create computer-aided diagnostic programs. A crucial component of computer-aided diagnostic testing is machine learning (ML). Body organs, for example, cannot be accurately identified using a simple equation. As a result, pattern recognition basically needs instance-based training. Pattern recognition and machine learning hold potential for enhancing the accuracy of disease approach and diagnosis in the biomedical field. They also honor the impartiality of the decision-making process. For the analysis of high-dimensional and multi-modal biomedical data, machine learning (ML) offers a credible method for creating better and automated algorithms. This survey paper provides a comparative analysis of different machine learning algorithms for the diagnosis of different diseases, including diabetes and heart disease. It focuses on the set of machine learning algorithms and methods used for decision-making and illness detection.**

***Index terms*—Naïve Bayes algorithm, Machine LearningAlgorithm, Artificial Intelligence**

## I. INTRODUCTION

Artificial intelligence can be thought through by the machine. AI increases the intelligence of machines. ML is a branch of AI research. According to several scholars, learning is necessary for the production of knowledge. The goal of machine learning is to create computer algorithms that can read data and utilize it to make their own decisions. The learning process begins with observation or data, such as references, firsthand experience, or instruction, so that we can look for trends in the data and make wise decisions in the future based on the examples we have. Allowing systems to learn on their own and adapt their behavior without human intervention or help is the main goal.

## II. LITERATURE SURVEY

The number of researchers who have worked on different machine learning algorithms for disease diagnosis is covered in this section. Researchers have recognized that machine-learning algorithms are effective in diagnosing a variety of illnesses. Diabetes and heart disease are the conditions that MLT detected in this survey article.

### Heart Disease

A platform for tracking and study was presented by Otoom [2]. This suggested tool monitors and diagnoses coronary artery disease. The Cleveland Heart Data Collection is where UCI is taken from. There are 304 cases and 77 features/attributes in this data collection. Fourteen qualities are used out of 76 attributes. Three algorithms—Bayes-Net, SVM, and FT—are used in two tests for detection. The WEKA tools are used to identify. After practicing with the Holdout test, the SVM approaches achieve an accuracy of 88.3 percent. In the Cross Validation test,

both SVM and Bayes-Net provide a precision of 83.8%. Accuracy of 81.5 percent is attained after employing FT. The best features of FT.7 are gathered using the Best First chosen algorithm, and cross-validation checks are employed for assessment. By applying the test to the seven best-selected features, Bayes Net obtained 84.5 percent accuracies, SVM provides 85.1 percent accuracies, and FT correctly classifies 84.6 percent. Vembandasamy [3] suggested that the Naïve-Bayes algorithm be used in a study to detect heart conditions. Included in Naïve-Bayes is Baye's theorem. The Naïve-Bayes have a strong presumption of freedom as a result. One of the top diabetic research institutes in Chennai provided the data that was used. The data collection includes 500 patients. Classification is carried out utilizing Weka as a method and 70% of Percentage Split. Precision with Naive Bayes is 86.419%. Tan [4] suggested a hybrid solution in which two machine learning algorithms, Genetics Algorithm (GA) and SVM, are successfully joined utilizing the wrapper method. LIBSVM and the WEKA data mining tool are employed in this investigation. Two data sets (heart disease and diabetes) will be retrieved for this investigation from the UC Irvine ML repository. The GA and SVM hybrid approach results in an accuracy of 84.07 percent for heart disease. The accuracy rate for collecting data on diabetes is 78.26 percent. Additionally, it is a binary classifier to generate the correct classifier, which has the advantages of being less over-fitting and noise-resistant, but it also has the disadvantages. For the classification of multiple classes, pairwise identification may be used. It operates slowly due to the high expense of calculation.

Iyer [5] performed a study to forecast diabetes disease using Naïve-Bayes and decision trees. When insulin is used excessively or is not produced enough, diseases develop. The data set used in this study is the diabetes data set from Pima India. The data mining application WEKA was used to conduct a number of studies. In this data collection, the percentage division (71:31) predicts more accurately than cross-verification. Using Percent Splitting and Cross-Verification J48 shows precision of 74.8698 percent and 76.9565 percent, respectively. Naïve-Bayes offers precisions of 79.5653 percent when employing PS. Algorithms show the maximum precision when % split checks are used. A study to predict type 2 diabetes in Naïve-Bayes has been proposed by Sarwar and Sharma [6]. There are three types of diabetes. The first type of diabetes will be Type 1. Type 2 diabetes is the second kind, followed by gestational diabetes. The development of insulin resistance leads to Type 2 diabetes. Due to variety, there are 416 cases in the data collection; the data are gathered from various sectors and communities in India. Models are created using MATLAB and a SQL server. Naive Bayes forecasts with 95% accuracy. Ephzibah has devised a method for diagnosing diabetes [7]. The GA and fuzzy logic are joined by the proposed model. This enhances classification accuracy and helps choose the best feature subsets. MLUCI Lab gathers a dataset for research, which consists of 768 instances and 7 attributes. To deploy, MATLAB is utilized. The genetic algorithms are used to choose only three positive traits. These three features are utilized by fuzzy logical classifiers, which offer 88 percent accuracy. The difference is around half of the initial cost.

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

## III. PROPOSED SYSTEM

The overview of our proposed system is shown in the below figure.

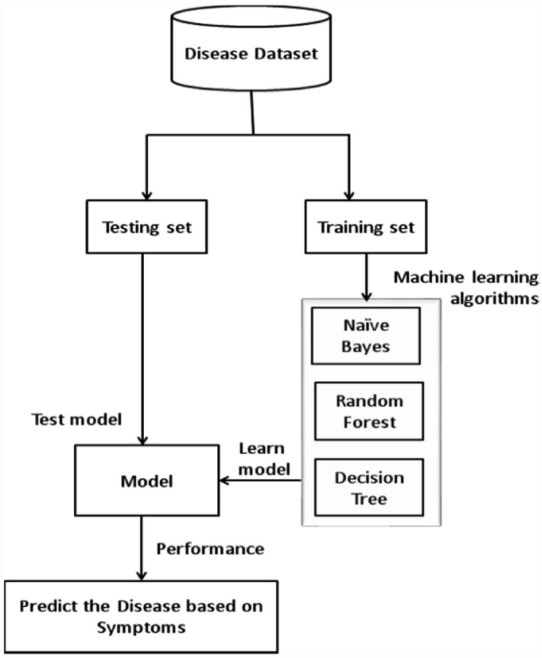


Fig. 1: System Overview

### *Implementation Modules*

**Load Dataset**

- In this phase, load the dataset into program and extract the data from the .csv file.
- This data can be analyzed and extract the best features to preprocess the data.
- Preprocess
- For the given data set, there are quite a few 'NA' values which are filtered in python. Furthermore, as the data set consists of numeric data, we used robust scaling, which is quite similar to normalization, but it instead uses the interquartile range whereas normalization is something which normalization shrinks the data in terms of 0 to 1.

**Split and Train and Test Model**

- In this module, the service provider split the Used dataset into train and test data of ratio 70 % and 30 % respectively. The 70% of the data is consider as train data which is used to train the model and 30% of the data is consider as test which is used to test the model.

**Prediction**

- In this module, the user enter the disease to predict the disease type. Random Forest and Naive Bayes from Sklearn for the disease prediction. The model has been pretrainedon a dataset of 4920 trials with 132 symptoms and 41 diseases.

### *Implementation Algorithms*

**Random forest**

- It creates a number of decision trees, each of which makes a prediction based on a portion of the data sample.
- The outcome that the greatest number of trees were able to produce is then regarded as the final forecast.
- A supervised learning algorithm called Random Forest employs the ensemble learning approach for regression and classification.
- Random forests are a bagging technique where trees operate in parallel without interacting with one another.
- During training, a Random Forest builds many decision trees, with the mean of the classes serving as each tree's prediction.

**Navie Bayes**

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

- Based on the Bayes theorem, the supervised learning technique known as the Naïve Bayes algorithm is used to solve classification issues.
- Text classification using a high-dimensional training dataset is its primary application.
- One of the simplest and most efficient classification techniques for creating quick machine learning models with rapid prediction capabilities is the Naïve Bayes Classifier.
- Because it is a probabilistic classifier, it makes predictions based on an object's likelihood.

## IV. RESULTS

The experimental outcomes of our implementing system are covered in this section. The figures depict how various diseases are predicted based on their symptoms.

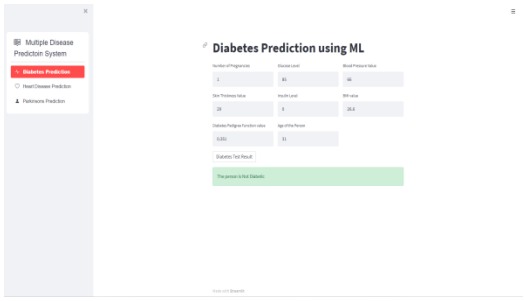


Fig. 1: Diabetic Prediction


Fig. 2: Parkinson 's disease Prediction

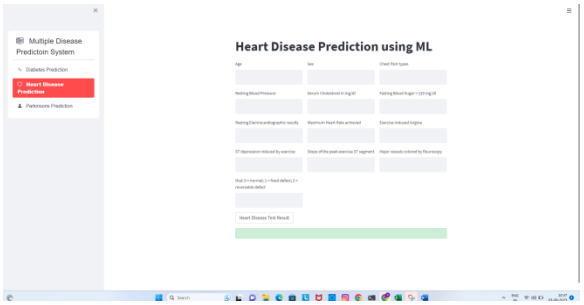


Fig. 3: Heart Disease Prediction

## SSV. CONCLUSION

Statistical prediction models that are unable to produce high-quality results have overtaken the evaluation field. Statistical models are ineffective at handling broad data points and missing values when it comes to preserving generalized information. The value of MLT originates from all of these causes. In many applications, ML plays a vital role, such as image recognition, data mining, processing of natural languages and diagnosis of diseases. In each of these areas, machine learning offers possible answers. This research explores several machine learning algorithms for diagnosing a range of illnesses, including diabetes and heart disease. Because they explain the characteristic in detail, the majority of models have produced great outcomes. According to earlier research, SVM improves performance for identifying heart disease by 94.60 percent. Naive Bayes is a diabetes condition that has been accurately diagnosed. It delivers 95 percent of the greatest categorization precision. The poll highlights the benefits and limitations of such algorithms. This survey report also presents a set of tools developed inside the AI community. These methods offer chances for a better decision-making process and are highly helpful for the examination of specific issues.

## REFERENCES

[1] Marshland, S. (2009) Machine Learnings an Algorithmic Perspectives. CRC Press, New Zealand, 6-7.

[2] Otoom et al., (2015) Effective Diagnosis and Monitoring of Heart Diseases. International Journal of Software Engineering and Its Application. 9, 143-156.

[3] Vembandasamy et al., (2015) Heart Disease Detection Using Naive Bayes Algorithms. IJISET-International Journal of Innovative Science, Engineering & Technology, 2, 441-444.

[4] Tan et al., (2009) A Hybrid Evolutionary Algorithm for Attribute Selections in Data Mining. Journal of Expert Systems with Application,36,8616-8630.https://doi.org/10.1016/j.eswa.2008.10.013.

[5] Iyer, A., Jeyalatha, S. and Sumbaly, R. (2015) Diagnosis of Diabetes Using Classification Mining Technique. International Journal of Data Mining & Knowledge Management Process (IJDKP), 5, 1-14.

[6] Sarwar, A. and Sharma, V. (2012) Intelligent Naïve Bayes Approaches to Diagnose Diabetes Type-2. Special Issues of International Journal of Computer Application (0975-8887) on Issues and Challenges in Networking, Intelligences and Computing Technologies-ICNICT 2012, 3, 14-16.

[7] Ephzibah, E.P. (2011) Cost Effective Approach on Feature Selection using Genetic Algorithm and Fuzzy Logics for Diabetes Diagnosis. International Journal on Soft Computing (IJSC), 2, 1-10. https://doi.org/10.5121/ijsc.2011.2101.

[8] Vijayarani, S. and Dhayanand, S. (2015) Liver Diseases Prediction using SVM and Naive Bayes Algorithm. International Journals of Science, Engineering and Technology Researches (IJSETR), 4, 816-820.

[9] Gulia, A et al., (2014) Liver Patients Classification Using Intelligent Technique. (IJCSIT) International Journal of Computer Science and Information Technology, 5, 5110-5115.

[10] Rajeswari, P. and Reena,G.S. (2019) Analysis of Liver Disorders Using Data Mining Algorithms. Global Journal of Computer Science and Technology, 10, 48-52.