PREDICTING DISEASES BASED ON SYMPTOMS WITH MACHINE LEARNING TECHNIQUES

[1]A Hemanth Kumar,[2]T. Hima Bindhu[3]B. Lakshmi Jeevana Pravallika, [4]V Anil Kumar,
[1]**Associate Professor** Department of Computer Science Engineering,
[234]**Assistant Professor**, Department of Computer Science Engineering,
AUDISANKARA COLLEGE OF ENGINEERING & TECHNOLOGY,
NH-16, By-Pass Road, Gudur, Tirupati Dist, Andhra Pradesh, -524101, India.

ABSTRACT—Computer Aided Diagnosis (CAD) is quickly evolving, diverse field of study in medical analysis. Significant efforts have been made in recent years to develop computer-aided diagnostic applications, as failures in medical diagnosing processes can result in medical therapies that are severely deceptive. Machine Learning (ML) is important in Computer Aided Diagnostic test. Object such as body-organs cannot be identified correctly after using an easy equation. Therefore, pattern recognition essentially requires training from instances. In the bio medical area, pattern detection and ML promises to improve the reliability of disease approach and detection. They also respect the dispassion of the method of decisions making. ML provides a respectable approach to make superior and automated algorithm for the study of high dimension and multi - modal bio medicals data. The relative study of various ML algorithm for the detection of various disease such as heart disease, diabetes disease is given in this survey paper. It calls focus on the collection of algorithms and techniques for ML used for disease detection and decision-making processes.

Index terms—Naïve Bayes algorithm, Machine Learning Algorithm, Artificial Intelligence.

INTRODUCTION

The machine can think through Artificial Intelligence. AI makes machines even more intelligent. The subfield of AI Research is ML. [2] Different researchers think that knowledge cannot be produced without learning's. The objective of ML is on designing computer algorithms that can read and use data to know for themselves.[5] In order to search for trends in data and make informed choices in the future based on the examples we have, the learning process starts with observation or data, such as references, direct experience, or guidance. The primary objective is to allow systems to learn automatically and change behavior according without human involvement or assistance.[8]

I. LITERATURE SURVEY

This segment discusses how many researchers have worked on various ML algorithms for disease diagnostic. [11]It has been acknowledged by researchers that machine-learning algorithms perform well for the diagnosis of various diseases. Diseases identified by MLT in this survey paper are heart and diabetes.

Heart Disease

Otoom [2] introduced a framework for research and tracking purposes. This proposed device detects and tracks coronary artery disease. UCI is extracted from the Cleveland heart data collection. This data set is made up of 304 cases and 77 features/attributes. Out of 76 features, 14 characteristics are used. For detection purposes, two experiments are performed with three algorithms: Bayes-Net, SVM, and FT. For identification, the WEKA tools is used. 88.3 percent accuracy is reached by using the SVM techniques after practicing with the Holdout test. The precision of 83.8 percent is given by both SVM and Bayes- Net in the Cross Validation test. After using FT, 81.5 percent accuracy is achieved. Using the Best First selected algorithm, FT.7 best characteristics are collected by and Cross- validation Checks are used for evaluation. Bayes Net achieved 84.5 percent accuracies by apply the test to 7 best selected feature, SVM offers 85.1 percent accuracies and FT properly classifies 84.6 percent. Vembandasamy [3] proposed a research was conducted using the Naïve-Bayes algorithm to identify heart diseases. In Naïve-Bayes, Baye's theorem is included. Therefore, the Naïve-Bayes have a strong presumption of freedom. The data collection used was collected from one of Chennai's leading diabetics research institutes. 500 patients are part of the data collection. By using 70 per cent of Percentage Split, Weka is used as a method and executes classifying.[4] Naive Bayes provides 86.419% precision. Tan [4] proposed in which two ML algorithms called Genetics Algorithm (GA) and SVM are effectively joins by using the wrapper method, the proposed hybrid strategy. In this study, LIBSVM and the WEKA data mine tool are used. For this analysis, two data sets (Diabetes disease, Heart disease) will be obtained from the UC Irvine ML repository.

84.07 percent precisions for heart disease is achieved after using the GA and SVM

hybrid strategy. 78.26 percent accuracies are reached for a diabetes data collection. And some of the benefits are that it is a binary classifier to create right classifier and less over-fitting, resilient to noise and the drawbacks are. It may use pair wise identification for the classification of multi- classes. [5]The cost of computation is high, so it works slowly.

Diabetes Disease

Iyer [5] is using decisions trees and Naïve- Bayes, they conducted a job to predict diabetes disease. Diseases arise when there is inadequate insulin production or there is excessive use of insulin. [10]The Pima India diabetes data set is the data set used in this work. Various experiments were carried out using the data mining tool WEKA. The percentage division (71:31) predicts better than cross-verification in this data collection. [8]By using Cross-verification and Percent Splitting Respectively, J48 indicates 74.8698 percent and 76.9565 percent precision. By using PS, Naïve-Bayes provides

79.5653 percent precisions. By using percent split checks, algorithms demonstrate the highest precision. arwar and Sharma [6] proposed a study to predict diabetes type 2 on Naïve-Bayes has been suggested. Diabetes has 3 forms. A diabetes Type-1 will be the first type. The diabetes type 2 second type, and the third type is gestational diabetes. A diabetes Type 2 results from the production of opposition to insulin. The data collection consists of 416 cases for the reason of varieties; data are collected in India from different societies of sector. For model creation, MATLAB with a servers SQL is used. Naive Bayes achieves 95% accurate forecasting. A diabetes diagnosis method has been developed by Ephzibah [7]. The GA and fuzzy logic are joined by the proposed model. This is used to select the finest subsets of features and also to improve classified accuracy. MLUCI Lab, which has 7 attributes and 768 instances, collects a dataset for study. For deployment, MATLAB is used. [6]Only three good characteristics are picked by using the genetics algorithms. The fuzzy logical classifiers use these three characteristics and provides 88 percent precision. Approximately 50 percent is less than the original expense.

II. PROPOSED SYSTEM

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

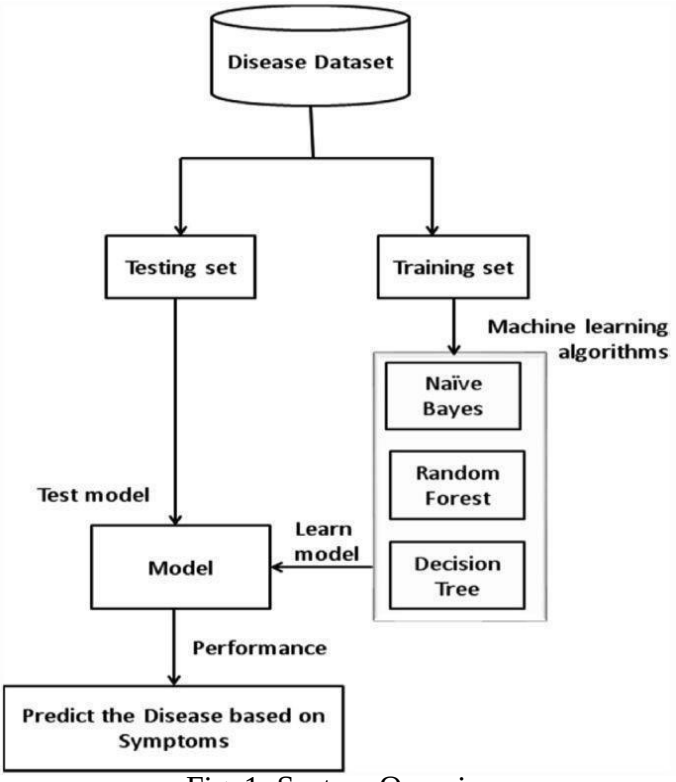

Fig. 1: System Overview

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 generates multi decision trees from which each decision tree uses a part of data sample and predicts the result.
- Then the result which was achieved by maximum number of trees is considered as the final prediction.
- Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression.
- Random forest is a bagging technique and the trees in random forests run in parallel without any interactions.
- A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.

Navie Bayes

- Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is mainly used in text classification that includes a high-dimensional training dataset.
- Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object

III. RESULTS

This section deals with experimental results of our implementing system. The figures shows the prediction of different diseases based on symptoms.

Fig. 1: Diabetic Prediction

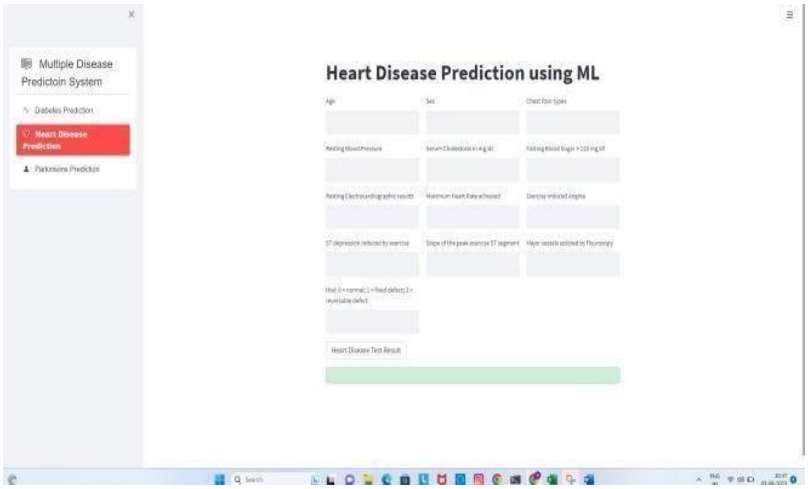

Fig. 2: Parkinson 's disease Prediction

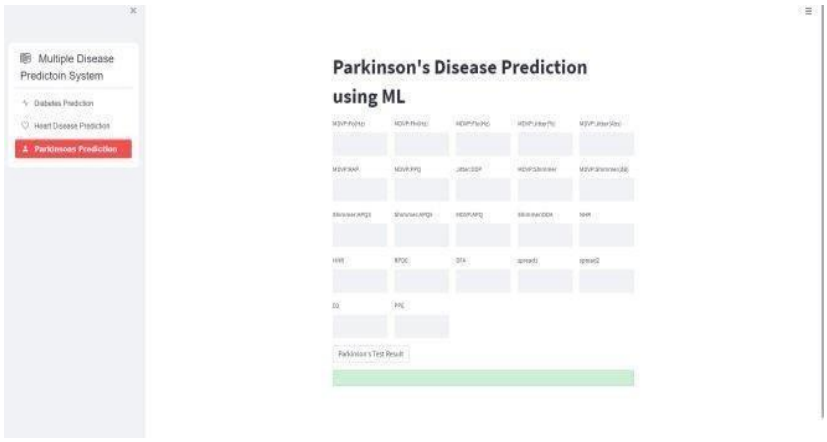

Fig. 3: Heart Disease Prediction

IV. CONCLUSION

The evaluation field has been flooded by statistical prediction models that are incapable of generating good quality outcomes. In maintaining generalized knowledge, statistical models are not efficient, coping with missing values and broad data points. The value of MLT stems 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. ML provides potential solutions in all these fields. This project discusses various techniques of ML for the diagnosis of various diseases such as heart, diabetes diseases. Most models have shown excellent results because they specifically describe the characteristic. It is noted from previous studies that SVM provides 94.60 percent improved performance for heart disease identification. Naive Bayes is a correctly diagnosed diabetes condition. It provides 95 percent of the highest classification precision. The survey shows the benefits and drawbacks of such algorithms. A suite of tools built in the AI community is also presented in this survey paper. These approaches are very useful for the analysis of certain problems and also provide opportunities for an improved decision-making process.

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] R. Seshadri, and N. Penchalaiah, "Noise analysis and detection based on RF energy duration in wireless LAN," Int. J. Distrib. Parallel Syst., Vol. 2, no. 5, Sep. 2011. pp. 57–66

[5] M Rajaiah, A Sudhakaraih, S V K Varma and P Venkatalakshmi (2015): Chemical and Soret effect on MHD free convective flow past an accelerated vertical plate in presence of inclined magnetic field through porous medium. i-manager's Journal on Mathematics, Vol. 4(1), pp. 32-39

[6] Mr. Ch. Dayakar, V. Chinnabbigari Srinivasulu, "Stock Portfolio Management System", Journal of Engineering Sciences, Vol 15, Issue 08, ISSN:0377-9254, 2024.

[7] Ephzibah, E.P. (2011) Cost Effective Approach on Feature Selection using Genetic Algorithm and Fuzzy Logics.