

Symptoms Based Disease Prediction Using Machine Learning Classifier**¹G Ramesh, ²S Suresh & ³Korrawath Ravi Naik**

¹Assistant professor, Department of CSM, CMR College of Engineering and Technology (CMRCET),
E- Mail: monty.ramesh21@gmail.com

²Assistant Professor, Department of IT, CMR College of Engineering and Technology (CMRCET),
E- Mail: - sureshse09@gmail.com

³Assistant professor, Department of CSM, CMR College of Engineering and Technology (CMRCET),
E- Mail: ravinaiikchauhan91@gmail.com

Abstract

Computer-Aided Diagnosis (CAD) is a 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 tests. An object such as body organs cannot be identified correctly after using an easy equation. Therefore, pattern recognition essentially requires training from instances. In the biomedical area, pattern detection and ML promise 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 making a superior and automated algorithm for the study of high dimension and multi-modal bio medicals data. The relative study of various ML algorithms for the detection of various diseases such as heart disease and diabetes disease is given in this survey paper. It calls focuses on the collection of algorithms and techniques for ML algorithms such as Naive Bayes Classifier, Random-Forest, Decision Tree, and Voting Classifier used for disease detection and decision-making processes.

Keywords: - CAD, ML Algorithms, Symptoms,

1. INTRODUCTION

Our medical care area every day gathers enormous information worried about patients including clinical assessment, imperative boundaries, examination reports, therapy subsequent meet-ups, drug choices, and so forth Yet, tragically it isn't examined and mine in a suitable manner. It is taken care of either in the record room as bunches of the paper sheet or devouring hard circle space. The experts similarly like investigators are hasty and stressed over this huge data. In government, just as in certain associations the information is handling mostly by analysts at proficient level. The improvement of mechanized structures and their precision will oversee us in future. It will

supportive in different illnesses the executives including viability of surgeries, clinical trials, drug, and the disclosure of connections among clinical and determination information to utilize Machine Learning systems. The medical care and clinical area are more needing data mining today. At the point when certain information mining techniques are utilized in a correct manner, significant data can be removed from enormous data set and which could guide the clinical professional to draw rapid choice and upgrade wellbeing administrations. The soul point is utilizing the grouping so that it can help doctor. Illnesses also good fitness associated issues such as intestinal sickness, Chickenpox, Migraine, Diabetes, Impetigo, Jaundice, dengue

and so forth, tend to critical impact on person's wellbeing and at times may likewise prompt passing whenever disregarded. The medical services industry can settle on a powerful dynamic by "mining" the vast information base they have for example by removing the secret examples also, connections in the data set. The machine learning models like Naive Bayes Classifier, Random-Forest, Decision Tree, and Voting Classifier or models can give a solution for the present circumstance. Subsequently, we have fostered a robotized framework that can find and separate secret information related with the infections from a historical (diseases-side effects) data set by the standard set of the individual Algorithms and models.

The main objective of the paper is

- To characterize the illnesses by utilizing different calculations like Random Forest, Naive Bayes, Decision Tree, and Voting Classifier.
- To track down the most affecting danger factors causing these illnesses.
- Comparison of different arrangement procedures and tracking down the best characterization strategy for the given information.
- To investigate the impact of change of one danger factor by another during the characterization (e.g., diabetes by hypertension, coronary illness, or smoking).

2. LITERATURE SURVEY

As per a description by McKinsey [1], half of Americans have at least one of the ongoing illnesses, and 80% of American clinical consideration charge is exhausted on persistent illness treatment. With the improvement of expectations for everyday comforts, the rate of persistent sickness is expanding. The United States has spent a normal of 2.7 trillion USD yearly on persistent illness treatment. This sum involves 18% of the whole yearly GDP of the United States. The medical services issue of ongoing infections is additionally very significant in numerous different nations. In China, persistent sicknesses are the primary driver of death, as per a Chinese report on sustenance and ongoing infections in 2015, 86.6% of passings are brought about by persistent sicknesses.

In this manner, it is fundamental to perform hazard evaluations for persistent illnesses. With the development in clinical information [2], gathering electronic wellbeing records (EHR) is progressively helpful [3]. Moreover, [4] first introduced a bio-inspired superior heterogeneous vehicular telematics worldview, with the end goal that the assortment of versatile clients' health related continuous large information can be accomplished with the arrangement of cutting edge heterogeneous vehicular organizations.

Chen et.al [5]– [7] suggested a clinical benefits system utilizing smart clothing for acceptable prosperity noticing. Qiu et al. [8] had inside and out examined the miscellaneous systems and attained the best results for cost reduction on the tree and primitive path cases for heterogeneous structures.

Patients' important genuine Data, results of various tests, and illness history are securely stored in the EHR, appealing us to find the more possible data-based driven approach mechanism for declining the costs of clinically relevant examinations.

Qiu et al. [9] described an effective flow surveying estimation for the tele-wellbeing cloud structure and arranged an information knowledge show for the PHR (Personal Health Record)- related spread system. Bates et al. [10] described six utilizes of huge amount of data in the field of clinical benefits. Qiu et al. [11] described a flawless huge information sharing calculation to deal with the muddle informational index in tele-health with cloud procedures.

Here one of the main approaches is to find most-danger patients with the help of the information, so that it can be utilized to minimize the clinical price since most-peril patient's routinely need to do have expensive clinical benefits besides, in the first paper proposing medical services digital actual framework [12], it inventively presented the idea of forecast based medical services applications, including wellbeing hazard evaluation. Figure using standard disorder danger models customarily incorporates an AI estimation (e.g., determined backslide and backslide examination, etc), and mainly an oversight learning computation by the utilization of planning information with names to set up the model [13], [14].

In the test dataset, patients can be broadly differentiated into social events of either more-peril or for the most part protected. These models are critical in clinical conditions and are by and large

considered [15], [16]. Regardless, these plans have the going with qualities and deformations. The instructive assortment is pretty much nothing, for patients and ailments with certain conditions [17], the attributes are picked through knowledge. Nonetheless, these pre-chosen qualities possibly not fulfill the progressions in the sickness and its influencing factors.

3. IMPLEMENTATION

System Model

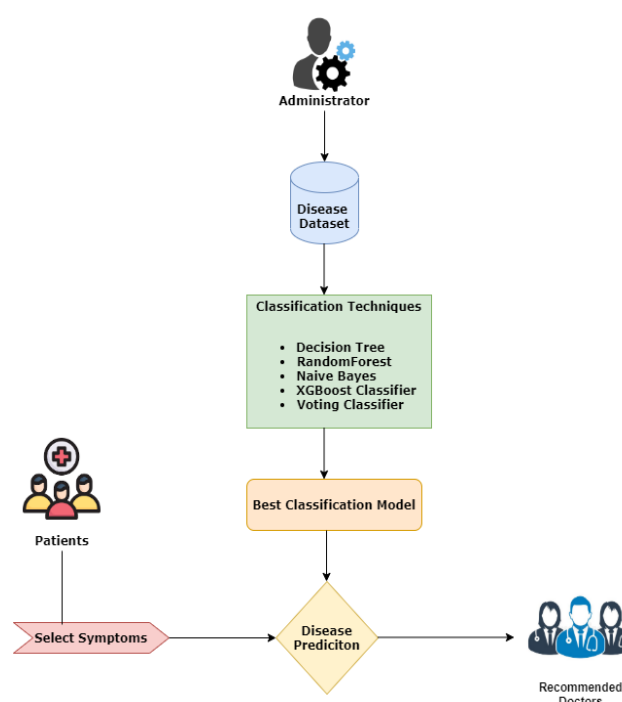


Figure.1 System Architecture

The figure.1, the patients and doctors will be registered with this system and the administrator will analyze the classification techniques with help of multiple disease datasets and determine the best model based on the highest accuracy. The patient will authenticate with this system and select the disease symptoms as input then this system will predict the diagnosis disease with the best classification model and this system will display

the recommended doctor's list based on the predicted disease.

Dataset Collection

In this research, I am using the Disease Prediction dataset which is derived from the Kaggle data collection. The dataset contains 4921 instances and 133 features or characteristics are treated as disease symptoms which are shown in figure.2.

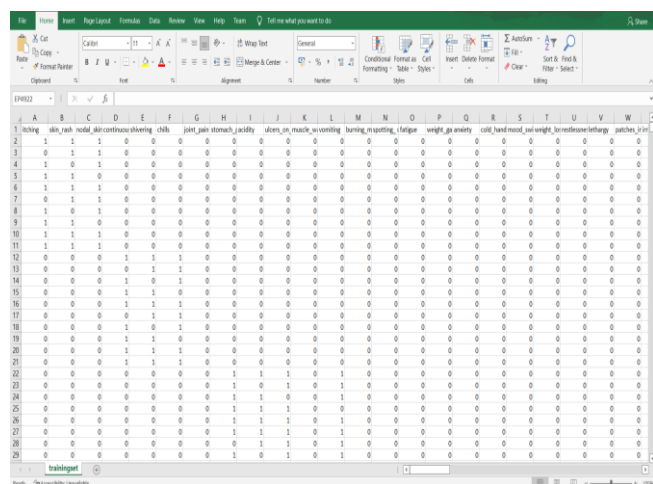


Figure.2. Disease Prediction Dataset

Data Pre-processing

In the data pre-processing stage, I have been loaded the dataset with help of the *panda's* library and separated the independent variables such as dataset features, and dependent variables like the target feature. So, here the features are treated as symptoms and the target feature is the predicted disease name.

Classifier Methods

RandomForest:

The RF classifier is a supervised machine learning algorithm and is mostly will use for regression and classification problems. The RF will predict the results based on multiple outcomes of a number of decision trees and which result is voter more time that one becomes as a predicted class. In this system, I used *RandomForestClassifier* which is

imported from *sklearn. ensemble* package for disease prediction. In the experimental results, the RF classifier is provided with 88 percent accuracy.

Naïve Bayes:

The Naive Bayes classifier also belongs to the supervised machine learning algorithm and it will process based on the Bayes theorem. The Bayes theorem will calculate the probability of all predicted diseases which is provided the highest probability that is predicted disease. I had used the *BernoulliNB* classifier which is imported from *sklearn. naive_bayes* package for disease prediction. In the experimental results, the NB classifier is provided 86.17 percent accuracy.

Decision Tree:

The DT classifier is a tree-based predicted classifier and it is also a supervised learning classifier used for resolving the classification problems. The DT classifier will predict the disease based on the IF-THEN methodology. While training the classifier with help of a dataset, the DT classifier will form the tree structure, so based on testing inputs, it will compare with all child nodes and it will reach the leaf node that is a predicted class. In this system, the *DecisionTreeClassifier* is used for disease prediction and this predefined classifier is imported from *sklearn. tree* package. Based on experimental results, the DT provided an accuracy of 82.18 percent.

Voting Classifier:

In this research, I have implemented a Voting classifier as my contribution to improving the prediction accuracy. It is a machine learning estimator because it will train with help of the

classifier models or estimators and predict based on the voting of each estimator output. It has two types of voting systems such as hard voting and soft voting. Here Hard voting will work based on the predicted output class and Soft voting will process based on the predicted output probability of the output class. According to experimental results, the voting classifier is performing the disease prediction with 95 percent. The predefined *VotingClassifier* is used for disease prediction and it is also imported from *sklearn. Ensemble* package.

Admin:

- The system will allow the admin to log in with a username and password.
- The system will allow the admin to approve the patients.
- The system will allow the admin to approve the doctors.
- The system will allow the admin to perform the machine learning evaluations.

Patient:

- The system will allow users to register with a phone number, name, username (username must be unique), and password.
- The system will allow the valid user to log in with his username and password.
- The system will allow the user to enter symptoms (n number of symptoms separated by commas).
- The system will allow the user to view the predicted disease.
- The system will allow the user to see the list of doctors along with their details (all the

attributes of doctors except username and password).

- The system will allow the user to select a doctor and book an appointment (while booking appointment patient details such as his name, gender, age, predicted disease should be sent to the doctor).
- The above-given details are sent to the doctor for appointment booking in the form of notifications.
- The system will allow users receive notification from doctors about the appointment details (scheduled time).
- The system will allow the user to rate the doctor.

Doctor:

- The system will allow doctors to log in with a username and password.
- The system will allow doctors to view all the appointment requests of users.
- The system will allow doctors to schedule the appointments.
- The system will allow the doctor to send notifications to the user on the scheduled timings.
- The ratings of the doctor will be calculated based on the average of all ratings provided by the user.
- The system will allow the user to view his (doctor) ratings.

EXPERIMENTAL RESULTS

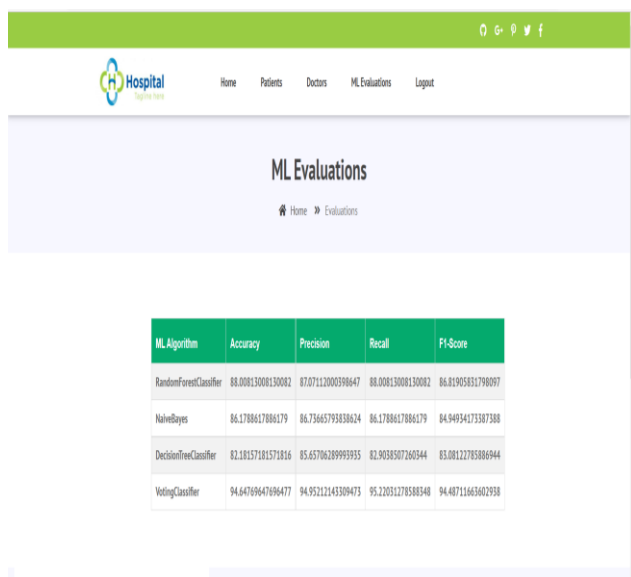


Figure.3 ML Evaluations

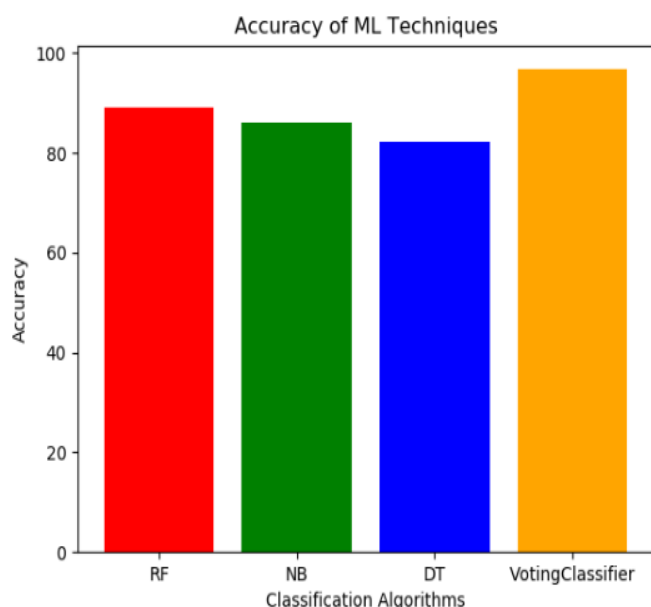


Figure.4 Accuracy of ML Techniques

4. 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 paper discusses various techniques of ML for the diagnosis of various diseases such as heart, and diabetes diseases. Most models have shown excellent results because they specifically describe the characteristic. It provides 95 percent of the highest classification precision. These approaches are very useful for the analysis of certain problems and also provide opportunities for an improved decision-making process.

5. REFERENCES

1. S. Mitra, S.K.Pal & Mitra , P., Data mining in soft computing framework: A survey, IEEE transactions on neural networks, 13(1), 314,2018.
2. Krzysztof J. Cios, G.William Moore, Uniqueness of medical data mining, Artificial Intelligence in Medicine 26, 1–24, 2017.
3. Parvez Ahmad, Saqib Qamar, Syed QasimAfser Rizvi, Techniques of Data Mining in Healthcare: A Review, International Journal of Computer Applications (0975 – 8887) Volume 120 – No.15, June 2017.
4. Hsinchun Chen, Sherrilynne, S. Fuller, Carol Friedman and William Hersh, Knowledge Management, Data Mining and text mining inmedical informatics.
5. V. krishnaiah, G. Narsimha, & N. Subhash Chandra, A study on clinical prediction using Data Mining techniques, International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) ISSN 2249-6831 Vol. 3, Issue 1, 239 248, March 2017.
6. Divya Tomar and Sonali Agarwal , A survey on data mining approaches for healthcare,



International Journal of Bio-Science and Bio-Technology Vol.No.5, pp. 241-266, 2017.7. Mohammed Abdul Khalid, Sateesh kumar Pradhan, G.N.Dash, F.A.Mazarbhuiya, A survey of data mining techniques on medical data for finding temporally frequent diseases”, International Journal of Advanced Research in Computer and Communication Engineering Vol.2, Issue 12, December 2018.

7. S.D.Gheware, A.S.Kejkar, S.M.Tondare, Data Mining: Task, Tools, Techniques and Applications, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 10, October 2017.

8. Yongjian Fu , Data Mining : Tasks, Techniques and Applications

<http://academic.csuohio.edu/fuy/Pub/pot97.pdf>

9. Pang-Ning Tan, Michael Steinbach, Vipin Kumar, "Introduction to Data Mining", Addison Wesley, 2017.

10. G. Beller, J. Nucl. Cardiol. “The rising cost of health care in the United States: is it making the United States globally noncompetitive?” vol. 15, no. 4, pp. 481-482, 2018.

11. Pang-Ning Tan, Michael Steinbach ,Vipin Kumar, "Introduction to Data Mining", Addison Wesley, 2016.

12. Gosain, A.; Kumar, A., "Analysis of health care data using different data mining techniques," Intelligent Agent & Multi-Agent Systems, 2017. IAMA 2009, International Conference on, vol. no., pp.1,6, 22-24 July 2018.

13. Dr. M.H.Dunham, “Data Mining, Introductory and Advanced Topics”, Prentice Hall, 2017. 14. A. S. Elmaghraby, et al. Data Mining from multimedia patient records. 6, 2017.

15. Nada Lavrac, BlažZupan, "Data Mining in Medicine" in Data Mining and Knowledge Discovery Handbook, 2018.

16. Soni J, Ansari U, Sharma D, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887), Volume 17– No.8, March 2018.

17. Naren Ramakrishnan, David Hanauer, Benjamin J. Keller, Mining Electronic Health Records, IEEE Computer 43(10): 77-81, 2018.