Symptom Based Disease Prediction Using Machine Learning

Ridham Sood, Virat Sharma

***Abstract:** The Disease Prediction Method uses predictive modeling to predict the user's disease based on the symptoms that the user offers as feedback to the system. Medical servicesare in desperate need to be advanced in order to make better choices about patient care and treatment options. In terms of machine learning, Healthcare enables humans to process large and complex medical databases, interpret them, and derive clinical insights. The machine analyzes the user's symptoms asinput and returns the disease's likelihood as an output. Implementing the Decision Tree, K Nearest Neighbor, Naïve Bayes and Random Forest allows for disease prediction. In thispaper, we attempt to integrate machine learning capabilities inhealthcare into a single framework. Instead of diagnosis, healthcare can be made smart by implementing disease prediction using machine learning predictive algorithms. Whenan early diagnosis of a disease is not possible, certain cases may arise. As a result, disease prediction can be applied effectively.This paper focuses primarily on the creation of a scheme, or what we would call an immediate medical provision, that would integrate symptoms obtained from multisensory devices as wellas other medical data and store it in a healthcare dataset. This Dataset would be analyze using machine learning algorithm with accuracy more than 90%.*

***Keywords:** Machine Learning Disease Prediction, Decision K Nearest Neighbor, Naïve Bayes, Random Forest.*

I. INTRODUCTION

Machine learning is the process of programming computers to improve their output based on previous data or examples. The study of computer systems that learn from data and experienceis known as machine learning. There are two tracks in the machine learning algorithm: training and testing. Prediction ofa disease based on the signs and medical history of the patient Machine learning has come a long way in the last few decades.

A portable health tracking system is proposed as a solution to the issue of advanced wearable devices. The main aim is to provide a wireless, low-cost, and user-friendly device that allows subjects to monitor clinical findings such as body temperature and heart rate, allowing doctors to regulate the subject's disease easily and rapidly from afar.

The primary goal is to use machine learning in healthcare to complement patient care and improve outcomes. Machine learning has simplified the process of accurately diagnosing and identifying various diseases. Predictive analysis using effective multiple machine learning algorithms aids in more accurate disease prediction and treatment of patients. Machine learning is now so pervasive that it is possible to use it many times a day without even realizing it. Machine learning algorithms, on the other hand, only work with structured data and have a long computation time since they store all of the data as a training dataset and use a complex calculation process.

Manuscript received on 25 July 2024 | Revised Manuscript received on 13 August 2024 | Manuscript Accepted on 15 September 2024 | Manuscript published on 30 September 2024.
*Correspondence Author (s)

Ridham Sood*, SRM Institute of Science and Technology, Chennai (Tamil Nadu), India. Email ID: rr8432@srmist.edu.in

Virat Sharma, SRM Institute of Science and Technology, Chennai (Tamil Nadu), India. Email ID: vj6767@srmist.edu.in

II. PROBLEM DEFINITION

The project's goal is to predict the disease by passing symptomsinto it. Traditional disease risk models typically use a machinelearning and supervised learning algorithm to train the models,which uses training data with labels.

EHR records patient statistics, test results, and disease history,allowing for the identification of possible data-centric strategies that lower the cost of medical case studies. Six applications of big data in healthcare are proposed by Bates et al. Diseases can be predicted by existing schemes, but not disease subtypes. It is unable to predict people's health.

III. PROBLEM SOLUTION

The proposed system of disease prediction using machine learning is that we have used many techniques and algorithms and all other various tools to build a system which predicts thedisease of the patient using the symptoms and by taking those symptoms we are comparing with the system's dataset that is previously available. By taking those datasets and comparing with the patient's disease we will predict the accurate percentage of the disease of the patient.

The dataset and symptoms go to the prediction model of the system where the data is pre-processed for the future referencesand then the feature selection is done by the user where he willenter the various symptoms. Then classification of those data isdone with help of various algorithms and techniques such as Decision Tree, KNN, Naïve Bayes and Random Forest.

IV. LITERARY SURVEY

A. Symptoms Based Disease Prediction Using Decision Treeand Electronic Health Record Analysis [1]

It has a number of attributes (symptoms) as well as classes (diseases). To train the model, we use this to build the training and testing sets. We obtain the user's symptoms and use the qualified model to predict the disease. In the other side, the medical record is compiled, which results in a review of the clinical report focusing on the most significant symptoms associated with a specific illness. This is used to increase the number of disease symptom pairs in the

Retrieval Number:100.1/ijpmh.G92340811922
DOI: 10.54105/ijpmh.G9234.04060924
Journal Website: www.ijpmh.latticescipub.com

Published By:
Lattice Science Publication (LSP)
© Copyright: All rights reserved.

dataset. This paper describes disease prediction using highly personalized trainingdata sets, as well as some related tasks such as scheduling appointments and locating the nearest health center.

B. Disease Prediction by Machine Learning Over Big Data from Healthcare Communities [2]

Using structured and unstructured data from hospitals, this paper proposes a new convolutional neural network-based multimodal disease risk prediction algorithm. In our knowledge, no current work in the field of medical big data analytics has centered on both data forms. As compared to other popular prediction algorithms, our proposed algorithm has a prediction accuracy of 94.8 percent and a convergence speed that is faster than the CNN-based unimodal disease risk prediction algorithm.

C. Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively [3]

This paper summarizes the findings of many studies in this area.Our proposed system seeks to bridge the divide between doctors and patients, assisting all groups in achieving their objectives. Using various Machine Learning algorithms, this framework supports multiple disease prediction. Many systems' current approach focuses solely on automating this method, which falls short of establishing user confidence in the system. By including a doctor’s advice in our system, we maintain consumer trust while also ensuring that the Doctors' business isnot harmed as a result of this system.

D. Multi Disease Prediction using Data Mining Techniques [4]

The accuracy of three data mining techniques is compared in this report. Along with high precision and recall metrics, the aim is to provide high accuracy. While these metrics are more commonly used in the field of information retrieval, we've included them here because they're relevant to other metrics including specificity and sensitivity. These metrics can be conveniently translated to true-positive (TP) and false-positive (FP) metrics using the uncertainty matric. For the prediction of various diseases, two separate data mining classification strategies were used, and their output was compared in order todetermine the best classifier. Building accurate andcomputationally effective classifiers for medical applications isa major challenge in data mining and machine learning.

E. Prediction of Heart Disease using Machine Learning Algorithms [5]

What we discovered is that during small datasets and in some other situations, decision trees often guide us to an incorrect solution; however, when we look at Nave Bayes results, we getmore reliable results with probabilities for all other possibilities, but decision trees can miss lead due to guidance to only one solution. Finally, we can assume that Nave Bayes is more reliable if the input data is cleaned and well maintained. While ID3 can clean itself, it cannot always produce accurate results, and similarly, Nave Bayes cannot always produce accurate results. We must consider the results of various algorithms, and if a prediction is made using all of their results, it will be accurate. However, we can use Nave Bayes to consider variables individually and combine algorithms such as Nave Bayes and K-means to achieve accuracy.

F. Review of Medical Disease Symptoms Prediction Using Data Mining Technique [6]

paper evaluates the performance of medical sickness predictionsupported data processing techniques. diagnosing of sickness knowledge like cancer, liver diseases, and heart attacks was classified into numerous categories by the classifier. With the utilization of 2 base classifiers, KNN and SVM, the SVM technique higher classified knowledge in compression as compared to the standard cluster ensemble method. we have a tendency to compare classification accuracy between customary algorithms and therefore the projected ensemble classification algorithms mistreatment a similar dataset. Additionally, to benchmark datasets from UCI repository, medical sickness datasets, and real-world datasets, we have a tendency to additionally examined the classification accuracy.

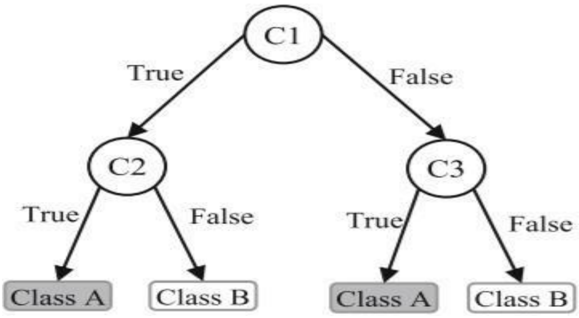
G. Disease Prediction from Various Symptoms Using Machine Learning [7]

Manuscript described a method of predicting disease based on the symptoms, age, and gender of a patient [8]. For the prediction of diseases, the weighted KNN model performed best at 93.5%with the factors listed above [9]. Almost all the machine learning models had good accuracy values [10]. Due to the dependence of some models on parameters, the models did not provide accurate predictions and their accuracy rates were quite low [11]. Once the disease has been predicted, we could easily manage the medicine resources required for treatment [12]. As a result, the disease would be treated less expensively and the recovery process would be enhanced.

V. ALGORITHMS

A. Decision Tree

One of the most well-known machine learning algorithms is the decision tree. A decision tree represents the decision logics forclassifying data items into a tree-like structure, i.e., tests and outcomes. A decision tree's nodes usually have several levels, with the root node being the first or top-most node. All internal nodes (those with at least one child) are input variableor attribute checks. The classification algorithm branches towards the appropriate child node based on the test result, andthe process of testing and branching repeats until the leaf nodeis reached.



[Fig.1: Decision Tree]

B. Naïve Bayes

The Bayes theorem is the basis for the Nave Bayes classification technique. This theorem can be used to explain the probability of an occurrence based on prior knowledge of the event's conditions. This classifier assumes that a feature in a class is not explicitly related to any other feature in the class,despite the fact that features in that class can be interdependent. The task of classifying a new entity into either class is used to demonstrate how the NB technique works.

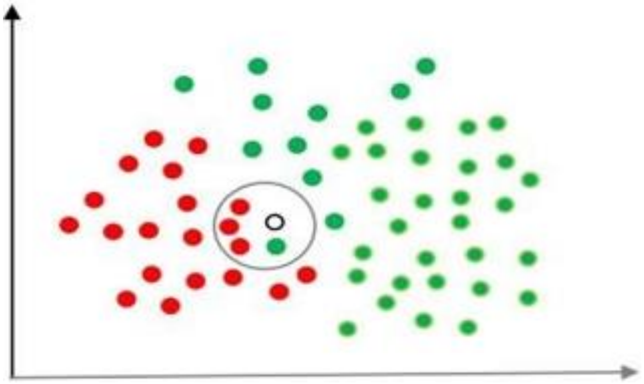
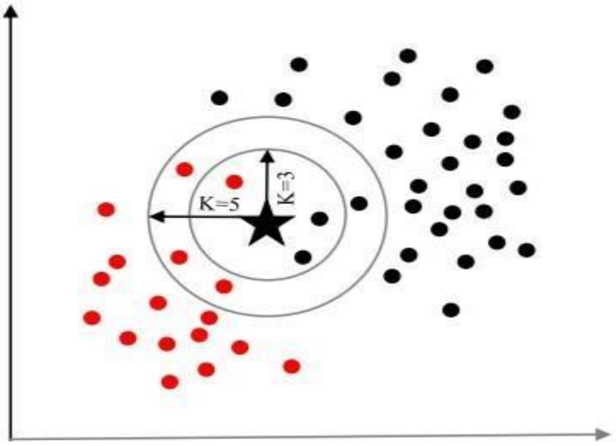


Fig. 2 Naïve Bayes

C. K-nearest Neighbor

One of the easiest and earliest classification algorithms is the K-nearest neighbor algorithm. It's like a simplified version of an NB classifier. The KNN algorithm, unlike the NB method, does not require the use of probability values. The number of nearest neighbors considered to take a ‘vote' is the ‘K' in the KNN algorithm. For the same sample item, different values for‘K' may result in different classification results. Depicts the KNN's classification process for a new object. When K=3, the new object (star) is classified as ‘black' but when K=5, it is classified as ‘red’.

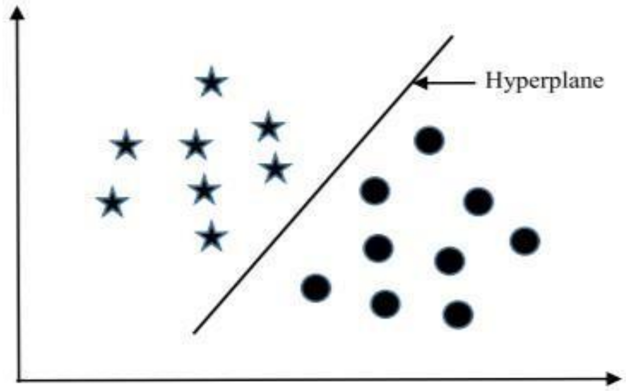


[Fig.3: K-Nearest Neighbor]

D. Random Forest

A Random Forest is a set of classifiers based on Decision Trees. A bootstrap sample of the data is used to construct each tree, which uses a candidate set of features chosen at random. For tree construction, it employs both bagging and random variable selection. After the forest has been developed, test cases are percolated down each tree, and the trees make their classpredictions. A random forest's error rate is determined by the intensity of each tree and the association between any two trees.

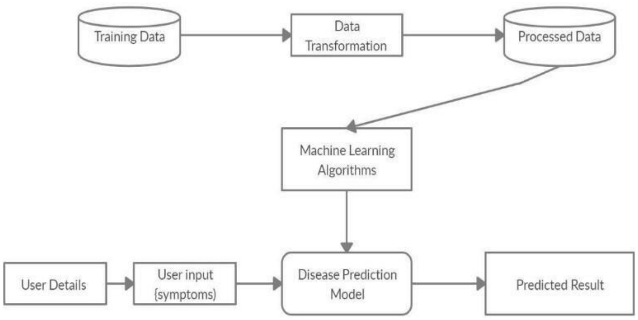
In a regression or classification query, it can be used to rank thevalue of variables in a natural way.



[Fig.4: Random Forest]

VI. ARCHITECTURE OF SYMPTOM BASED DISEASE PREDICTION USING MACHINE LEARNING

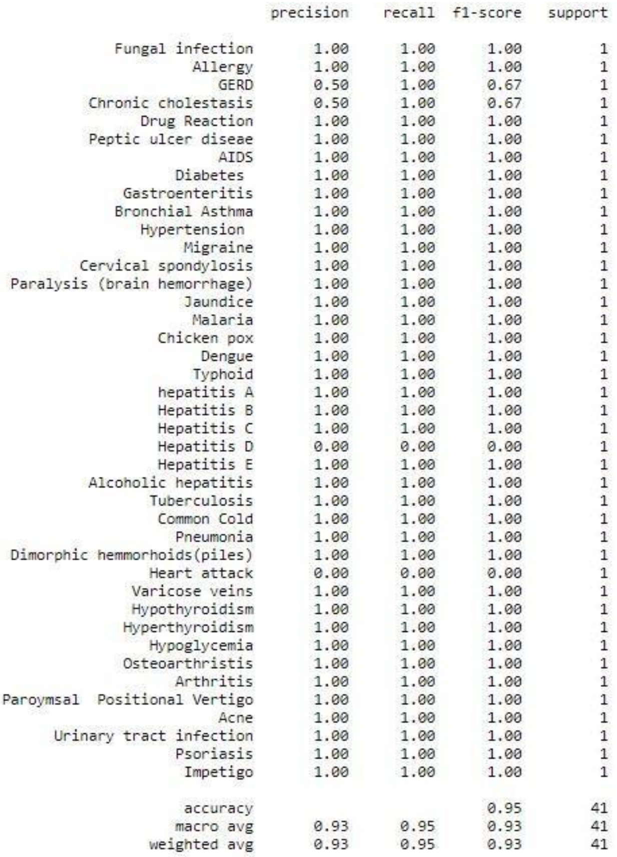
Machine learning-based disease prediction predicts the occurrence of a disease for a user based on different symptomsand the knowledge the user provides through the symptoms. The architecture of the system for disease prediction using machine learning consists of various datasets from which we can compare and predict the user's symptoms, after which the datasets are transformed into smaller sets and classified using classification algorithms, and finally the classified data is processed into machine learning technologies. The method then integrates and compares the above information and total processed data in the prediction model, and ultimately predictsthe disease.



[Fig.5: Architecture]

VII. RESULT

Higher precision can be achieved with the proposed method. Based on the proposed algorithm, we not only use structured data, but also the patient's text data. On the datasets, we checkedour algorithms. On the dataset, we achieved 95 percent accuracy for Decision Tree, Nave Bayes, and Random Forest, and 92.6 percent accuracy for K-nearest neighbor. The classification report mentioned below shows the accuracy of our algorithm for each disease, as well as the overall accuracy of our model.



```
                              precision    recall  f1-score   support

             Fungal infection      1.00      1.00      1.00         1
                      Allergy      1.00      1.00      1.00         1
                         GERD      0.50      1.00      0.67         1
          Chronic cholestasis      0.50      1.00      0.67         1
                Drug Reaction      1.00      1.00      1.00         1
          Peptic ulcer diseae      1.00      1.00      1.00         1
                         AIDS      1.00      1.00      1.00         1
                    Diabetes       1.00      1.00      1.00         1
              Gastroenteritis      1.00      1.00      1.00         1
             Bronchial Asthma      1.00      1.00      1.00         1
                Hypertension       1.00      1.00      1.00         1
                     Migraine      1.00      1.00      1.00         1
         Cervical spondylosis      1.00      1.00      1.00         1
 Paralysis (brain hemorrhage)      1.00      1.00      1.00         1
                     Jaundice      1.00      1.00      1.00         1
                      Malaria      1.00      1.00      1.00         1
                  Chicken pox      1.00      1.00      1.00         1
                       Dengue      1.00      1.00      1.00         1
                      Typhoid      1.00      1.00      1.00         1
                  hepatitis A      1.00      1.00      1.00         1
                  Hepatitis B      1.00      1.00      1.00         1
                  Hepatitis C      1.00      1.00      1.00         1
                  Hepatitis D      0.00      0.00      0.00         1
                  Hepatitis E      1.00      1.00      1.00         1
          Alcoholic hepatitis      1.00      1.00      1.00         1
                 Tuberculosis      1.00      1.00      1.00         1
                  Common Cold      1.00      1.00      1.00         1
                    Pneumonia      1.00      1.00      1.00         1
 Dimorphic hemmorhoids(piles)      1.00      1.00      1.00         1
                 Heart attack      0.00      0.00      0.00         1
               Varicose veins      1.00      1.00      1.00         1
               Hypothyroidism      1.00      1.00      1.00         1
              Hyperthyroidism      1.00      1.00      1.00         1
                 Hypoglycemia      1.00      1.00      1.00         1
              Osteoarthristis      1.00      1.00      1.00         1
                    Arthritis      1.00      1.00      1.00         1
Paroymsal  Positional Vertigo      1.00      1.00      1.00         1
                         Acne      1.00      1.00      1.00         1
      Urinary tract infection      1.00      1.00      1.00         1
                    Psoriasis      1.00      1.00      1.00         1
                     Impetigo      1.00      1.00      1.00         1

                     accuracy                          0.95        41
                    macro avg      0.93      0.95      0.93        41
                 weighted avg      0.93      0.95      0.93        41
```

[Fig.6: Classification Report]

VIII. CONCLUSION

An architecture diagram is a graphical representation of a collection of concepts that make up an architecture, such as itsvalues, elements, and components. The diagram depicts the machine software in the context of a system description.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been sponsored or funded by any organization or agency. The independence of this research is a crucial factor in affirming its impartiality, as it has been conducted without any external sway.
- **Ethical Approval and Consent to Participate:** The data provided in this article is exempt from the requirement for ethical approval or participant consent.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Authors Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCE

1. S Radhika, S Ramiya Shree, V Rukhmani Divyadharsini and A Ranjitha, (2020). Symptoms Based Disease Prediction Using Decision Tree and Electronic Health Record Analysis, EuropeanJournal of Molecular & Clinical Medicine. https://ejmcm.com/uploads/paper/71d1389ad5bd34bb9f80901d1343506d.pdf
2. Chen, Yixue Hao, Kai Hwang, LuWang, Lin Wang p, (2017). Disease Prediction by Machine Learning Over Big Data From Healthcare Communities, Min IEEE. DOI: https://doi.org/10.1109/ACCESS.2017.2694446
3. Rudra A. Godse, Smita S. Gunjal,Karan A. Jagtap, Neha S. Mahamuni, Prof. Suchita, (2019). Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively Wankhade IJARCCE. https://ijarcce.com/wp-content/uploads/2020/01/IJARCCE.2019.81210.pdf
4. K.Gomathi, Dr. D. Shanmuga Priyaa, (2016). Multi Disease Prediction using Data Mining Techniques, Research Gate. https://www.researchgate.net/publication/319851535_Multi_Disease_Prediction_using_Data_Mining_Techniques
5. Rajesh N, T Maneesha, Shaik Hafeez, Hari Krishna, (2018). Prediction of Heart Disease Using Machine Learning Algorithms, ResearchGate. DOI: https://doi.org/10.1109/ICIICT1.2019.8741465
6. Rahul Deo Sah1, Dr. Jitendra Sheetalani, (2017). Review of Medical Disease Symptoms Prediction Using Data Mining Technique, IOSR-JCE. DOI: https://doi.org/10.9790/0661-1903015970
7. Rinkal Keniya, Aman Khakharia, Vruddhi Shah, Vrushabh Gada, Ruchi Manjalkar, TirthThaker, Mahesh Warang, Ninad Mehendale, (2018). Disease prediction from various symptoms using machine learning, SSRN. DOI: http://dx.doi.org/10.2139/ssrn.3661426
8. Sharma, P., & Site, S. (2022). A Comprehensive Study on Different Machine Learning Techniques to Predict Heart Disease. In Indian Journal of Artificial Intelligence and Neural Networking (Vol. 2, Issue 3, pp. 1–7). DOI: https://doi.org/10.54105/ijainn.c1046.042322
9. Bhanuteja, T., Kumar, K. V. N., Poornachand, K. S., Ashish, C., & Anudeep, P. (2021). Symptoms Based Multiple Disease Prediction Model using Machine Learning Approach. In International Journal of Innovative Technology and Exploring Engineering (Vol. 10, Issue 9, pp. 67–72). DOI: https://doi.org/10.35940/ijitee.i9364.0710921
10. Tamilarasi, Dr. A., Karthick, T. J., R. Dharani, & S. Jeevitha. (2023). Eye Disease Prediction Among Corporate Employees using Machine Learning Techniques. In International Journal of Emerging Science and Engineering (Vol. 11, Issue 10, pp. 1–5). DOI: https://doi.org/10.35940/ijese.c7895.09111023
11. Razia, S., Babu, J. C., Baradwaj, K. H., Abhinay, K. S. S. R., & M, A. (2019). Heart Disease Prediction using Machine Learning Techniques. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 8, Issue 4, pp. 10316–10320). DOI: https://doi.org/10.35940/ijrte.d4537.118419
12. Dubey, S. K., Sinha, Dr. S., & Jain, Dr. A. (2023). Heart Disease Prediction Classification using Machine Learning. In International Journal of Inventive Engineering and Sciences (Vol. 10, Issue 11, pp. 1–6). DOI: https://doi.org/10.35940/ijies.b4321.11101123

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Lattice Science Publication (LSP)/ journal and/ or the editor(s). The Lattice Science Publication (LSP)/ journal and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.