Prognosticare: An Advanced Prediction and Recommendation Algorithm for Healthcare

Kruti Shah
Computer Engineering
D J. Sanghvi College of Engineering
Mumbai, India
krutishah957@gmail.com

Preksha Patel
Computer Engineering
D J. Sanghvi College of Engineering
Mumbai, India
patelpreksha24@gmail.com

Khushi Jobanputra
Computer Engineering
D J. Sanghvi College of Engineering
Mumbai, India
khushidjobanputra77@gmail.com

***Abstract*—Healthcare is a fundamental human right crucial for societal well-being and economic prosperity. However, countries like India face significant challenges in providing adequate healthcare due to limited funding and shortages of healthcare professionals. In this context, the rapid growth of healthcare data presents both opportunities and challenges. Our research addresses this by proposing ”Prognosticare,” an innovative prediction and recommendation algorithm for healthcare. By harnessing the power of machine learning, Prognosticare aims to leverage symptom-based data to personalize healthcare solutions, thereby improving healthcare delivery and patient outcomes.**

***Index Terms*—machine learning, decision tree classifier, random forest classifier, naive Bayes classifier, disease prediction, disease classification, prognostic recommendations.**

I. Introduction

Healthcare is a fundamental human right essential for fulfilling basic needs and enhancing quality of life. It plays a critical role in the overall development of a country, as a healthy population contributes to economic growth and prosperity. However, in countries like India, the healthcare sector has been inadequately funded, leading to neglect and challenges in meeting the needs of its vast population.

India's healthcare expenditure, at 1.4 percent of GDP, lags behind neighboring countries like Sri Lanka and Nepal. With just one doctor for every 1668 people, the country faces significant shortages in healthcare professionals. Despite these challenges, the healthcare sector has experienced exponential growth, generating vast amounts of data daily, including electronic health records (EHRs).

This abundance of data presents a major challenge for physicians and data analysts to convert into actionable knowledge for effective prediction and healthcare delivery. Our research addresses this challenge by proposing a symptom-based prediction system using machine learning techniques. This system aims to personalize healthcare solutions such as diet, medication, precautions, and workouts based on individual symptoms reported by users.

Our research presents a groundbreaking approach to personalized healthcare through a symptom-based prediction system. Utilizing machine learning techniques exclusively, our system tailors healthcare solutions such as diet, medication, precautions, and workouts based on individual symptoms reported by users. This paper discusses the practical implications of our system while detailing the process of evaluating three algorithms to select the most effective approach for prediction. By showcasing its potential to revolutionize healthcare delivery and improve patient outcomes, we highlight the significance of our algorithm selection methodology in advancing personalized healthcare.

II. Literature Review

[2]The paper evaluates different machine learning algorithms on Breast Cancer, Diabetes, and Heart Disease datasets. It preprocesses the data, selects significant attributes, and compares prediction accuracy. AdaBoost excels for Breast Cancer, Support Vector Machine for Diabetes, and Logistic Regression for Heart Disease. The paper concludes by comparing proposed models with existing methods, showing superior performance.

[1]The paper describes the development of a medical chatbot using natural language processing and machine learning. Trained on a dataset of symptoms and diseases, users can interact with the chatbot to discuss health issues. The chatbot identifies symptoms, predicts diseases using the K-nearest neighbor algorithm, and recommends treatment. Experimental results demonstrate the chatbot's accuracy in predicting diseases based on symptoms. It provides reliable information, enabling users to monitor their health and take necessary actions. Cost-effective, accessible, and user-friendly, the chatbot serves as a valuable tool for healthcare management.

[3]This research introduces a novel approach to predicting disease risks based on medical history, using collaborative filtering techniques. It adapts methods from marketing to analyze patient profiles and forecast future health risks. Key contributions include applying collaborative filtering to medicine, incorporating significance testing, developing a time-sensitive system, analyzing performance trends, and providing case studies to illustrate the system's benefits.

III. METHODOLOGY

A. Datasets Used

In our research, we integrated various datasets, including those on diets, symptoms, precautions, workouts, and medications, to develop and evaluate our prediction system. Additionally, we incorporated datasets related to symptom profiles, precautionary measures, workout routines, and medication regimens from reputable sources such as

https://www.kaggle.com/datasets/choongqianzheng/disease-and-symptoms-dataset and https://www.kaggle.com/datasets-/itachi9604/disease-symptom-description-dataset. Notably, our symptoms dataset comprised 4919 rows, facilitating robust analysis and personalized recommendations. These datasets collectively contribute to the training and validation of our system, enabling tailored recommendations to meet individual needs.

B. Model Overview

In the realm of predictive modeling for healthcare, various algorithms play pivotal roles in diagnosing diseases based on symptoms and other medical data. Among these, three notable classifiers stand out: Decision Tree Classifier, Random Forest Classifier, and Naive Bayes Classifier. Each of these classifiers offers distinct advantages and is suited to different types of datasets and prediction tasks.

Decision tree Classifier

- A decision tree is a supervised learning approach used for characterizing and regressing data. Its basic structure represents data distribution in a tree-like fashion. The process begins by calculating the entropy of each attribute, and then the root is split into sub-trees or branches based on maximum information gain and predefined rules. This recursive process continues with the attributes until leaf nodes are reached, providing the final result.

  $$Entropy\,(X) = -\sum_{i=1}^{n} p\,(x_i)\,log_b p(x_i)$$

  In the above Equation , the class entropy is computed, where n represents the set of classes in X, X denotes the current dataset being processed, and p(x) is the ratio of the number of elements in class n to the number of elements in set X.

  Information Gain = class entropy- entropy attributes

  The information gain, as shown in the above Equation, is calculated by subtracting the entropy of each branch from the class entropy. Decision trees accommodate both categorical and continuous data, typically represented as 0 or 1, or yes or no. They offer high accuracy by considering every attribute and analyzing them with a tree-like structure. Additionally, decision tree models are easy to understand, and rules can be generated effortlessly. However, this algorithm may suffer from overfitting, especially when the tree has many branches, making it challenging to compute and interpret.

  The Decision Tree Classifier constructs a tree structure where each node represents a decision based on a specific attribute, such as symptoms. Branches represent possible outcomes of that decision, ultimately leading to the prediction of the target variable, such as disease prediction in healthcare applications. Decision nodes split the data into sub-nodes based on attribute values, while leaf nodes signify the final classification decision.

Random Forest Classifier

- The Random Forest Classifier is a widely used machine learning algorithm known for its robust performance. It addresses the overfitting issues commonly associated with decision trees by constructing multiple decision trees from randomly selected subsets of the training data. The final prediction is determined by aggregating the outcomes from these individual trees. Random forest operates as an ensemble learning method, combining predictions from various decision tree classifiers trained on different subsets of the data.

Naive Bayes Classifier

- Naïve Bayes is a supervised learning approach used for probabilistic classification based on Bayes' theorem. It assumes that features within a class are independent of each other, simplifying the computation. In the context of disease prediction, this independence assumption allows the algorithm to handle missing values effectively and provides an advantage when dealing with large datasets. Bayes' theorem, represented as:

  $$P\,(Q|R) = \frac{P\,(R|Q)\,P(Q)}{P(R)}$$

  The Naïve Bayes Classifier calculates the probability of a class given a set of features by assuming independence among features. This approach, while simplistic, offers efficiency and scalability, particularly in handling high-dimensional datasets. Its ability to quickly train and classify makes it a valuable tool in disease prediction and other classification tasks.

C. Proposed System Architecture

Beginning with disease datasets, a training set is established for machine learning algorithms. Simultaneously, a separate testing set is created for performance evaluation. Three machine learning algorithms, Naive Bayes, Random Forest, and Decision Tree, are employed for disease prediction. Each algorithm learns from the training set to develop a predictive model. The trained models are then tested using the testing set to evaluate their performance. Based on the performance assessment, the accuracy and effectiveness of each model are determined. Subsequently, the trained models are utilized to predict diseases based on symptoms. Finally,

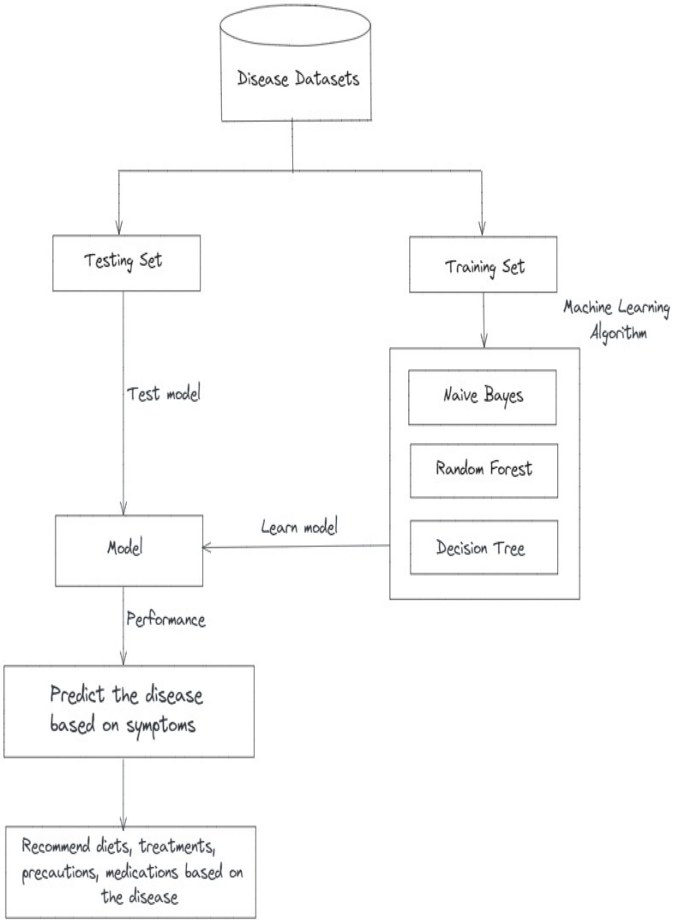

Disease Datasets
Testing Set
Training Set
Machine Learning Algorithm
Test model
Naive Bayes
Random Forest
Decision Tree
Learn model
Model
Performance
Predict the disease based on symptoms
Recommend diets, treatments, precautions, medications based on the disease

Fig. 1. Block Diagram of Proposed System Architecture

recommendations including diets, treatments, precautions, and medications are provided based on the predicted diseases, aiding in effective disease management and treatment.

IV. RESULTS AND EXPERIMENTS

The model was trained on medical record of 4920 patients. From the below table, we can infer that all the three algorithm shows excellent result but Naive Bayes performs the best and achieve the highest accuracy of 98.12 percentage. The training accuracy has been described in Table I. As we can see that the efficiency of training is higher in Naïve Based classifier. It is because it overcomes the problem of overfitting, which is common in the case of Decision Tree and Random forest classifiers.

TABLE I
TRAINING ACCURACY

Algorithm Used	Accuracy Score
Decision Tree	0.9763
Random Forest	0.9763
Naive Bayes	0.9812

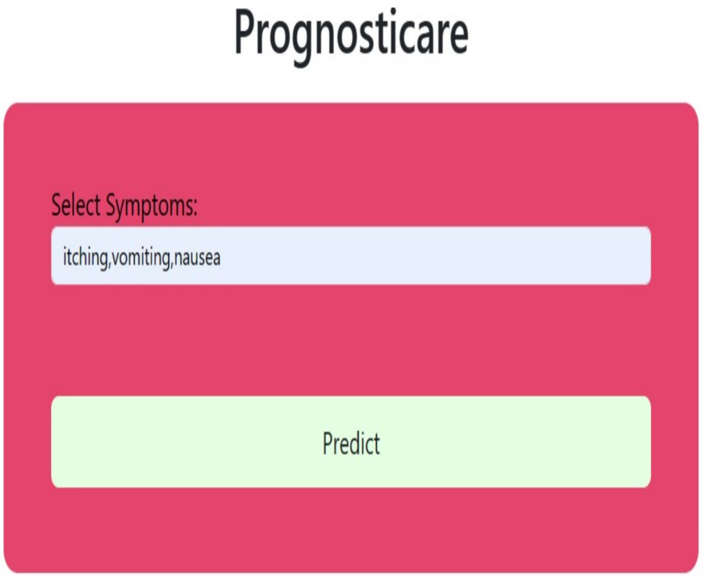
Once all these classifiers are trained, now it is ready for testing the results over any new disease. The model was tested on 41 new patient records. The accuracy score, number of diseases correctly classifier and wrongly classifier are mentioned in Table 6. Naïve Bayes classifier giving the highest score with 97% accuracy.

TABLE II
ACCURACY AND CONFUSION MATRIX

Algorithm	Accuracy Score	Correctly Classified	Wrong Classified
Decision Tree	0.95	39	2
Random Forest	0.95	39	2
Naive Bayes	0.97	40	1

A. Testing

Our system incorporates a user-friendly website interface, allowing individuals to conveniently input their symptoms and receive personalized predictions and recommendations. Figure 2 illustrates the intuitive symptom selection interface, enhancing accessibility and usability for users seeking healthcare guidance.

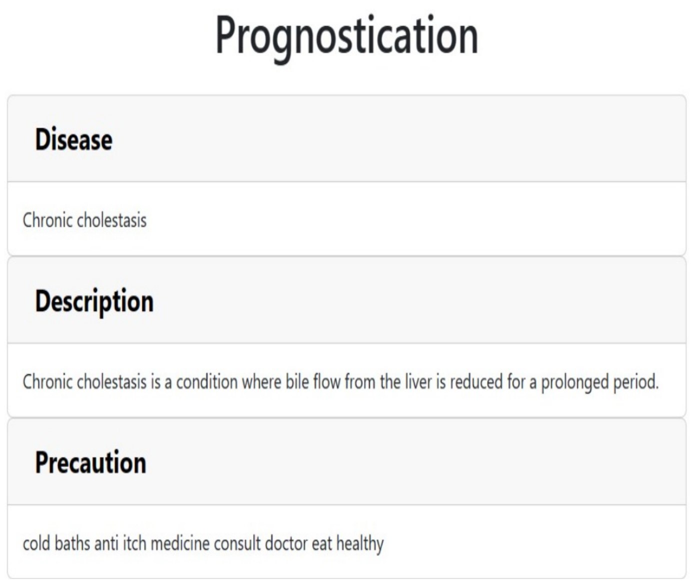


Select Symptoms:

itching,vomiting,nausea

Fig. 2. Selecting symptoms as per user requirements

Our platform facilitates informed decision-making by presenting users with comprehensive information about the predicted disease, including associated symptoms and potential complications. By delivering tailored recommendations for precautionary measures, treatments, medications, and dietary adjustments, we empower individuals to take proactive steps towards optimizing their health and well-being, fostering a culture of self-care and resilience.



Prognostication

Disease

Chronic cholestasis

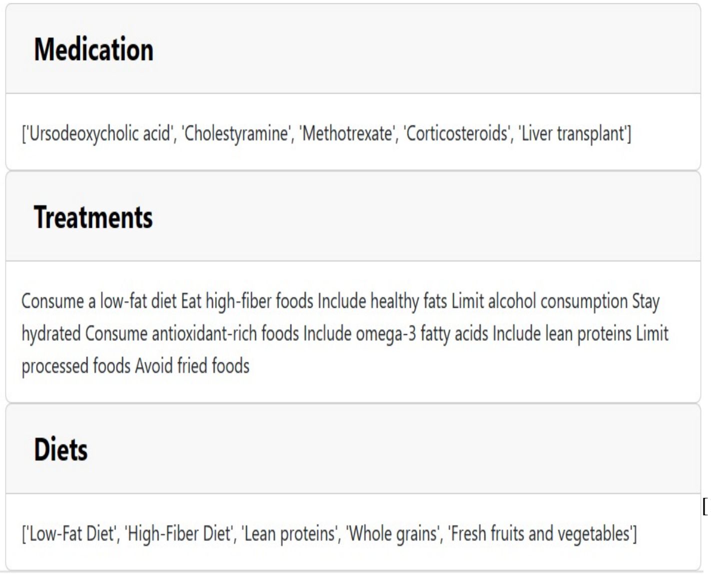
Description

Chronic cholestasis is a condition where bile flow from the liver is reduced for a prolonged period.

Precaution

cold baths anti itch medicine consult doctor eat healthy

Fig. 3. Displaying Prognosis as per user symptoms



Medication

['Ursodeoxycholic acid', 'Cholestyramine', 'Methotrexate', 'Corticosteroids', 'Liver transplant']

Treatments

Consume a low-fat diet Eat high-fiber foods Include healthy fats Limit alcohol consumption Stay hydrated Consume antioxidant-rich foods Include omega-3 fatty acids Include lean proteins Limit processed foods Avoid fried foods

Diets

['Low-Fat Diet', 'High-Fiber Diet', 'Lean proteins', 'Whole grains', 'Fresh fruits and vegetables']

Fig. 4. Displaying Prognosis as per user symptoms

V. CONCLUSION

In conclusion, our study introduces an innovative approach to personalized healthcare through a symptom-based prediction system. Leveraging machine learning techniques, we have developed a system capable of tailoring healthcare solutions such as diet, medication, precautions, and workouts based on individual symptoms reported by users.

Through a comprehensive evaluation of three prominent classifiers—Decision Tree, Random Forest, and Naive Bayes—we have demonstrated the effectiveness of each algorithm in predicting diseases based on symptoms. Our results indicate that while all three classifiers exhibit commendable performance, Naive Bayes outperforms the others with the highest accuracy score of 98.1%. This superior performance can be attributed to Naive Bayes' ability to mitigate overfitting issues commonly encountered in Decision Tree and Random Forest classifiers.

Furthermore, our testing phase on new patient records reaffirms the robustness of our prediction system, with Naive Bayes achieving an impressive accuracy score of 97

In summary, our research showcases the potential of machine learning in transforming healthcare delivery by providing personalized solutions tailored to individual needs. By accurately predicting diseases based on symptoms, our system empowers healthcare providers and patients alike to make informed decisions, leading to improved patient outcomes and overall well-being.

REFERENCES

[1] Rohit Binu Mathew; Sandra Varghese; Sera Elsa Joy; Swanthana Susan Alex, "Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning" 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)

[2] P. S. Kohli and S. Arora, "Application of Machine Learning in Disease Prediction," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-4, doi: 10.1109/CCAA.2018.8777449.

[3] T. M. Mitchell, "Decision Tree Learning," Machine Learning. pp. 52–80, 1997.

[4] Davis, D.A., Chawla, N.V., Christakis, N.A. and Barabási, A.L., 2010. Time to CARE: a collaborative engine for practical disease prediction. Data Mining and Knowledge Discovery, 20, pp.388-415.

[5] Davis, D.A., Chawla, N.V., Christakis, N.A. and Barabási, A.L., 2010. Time to CARE: a collaborative engine for practical disease prediction. Data Mining and Knowledge Discovery, 20, pp.388-415.

[6] Hamet, P. and Tremblay, J., 2017. Artificial intelligence in medicine. Metabolism, 69, pp.S36S40.

[7] Abhishek ,Amit Bindal, Dharminder Yadav, Abhishek Bindal, Dr. Amit Yadav, Dharminder. (2022). Medicine recommender system: A machine learning approach. 50015.

[8] Goyal, V.A., Parmar, D.J., Joshi, N.I. and Champanerkar, K., 2020. Medicine recommendation system. Medicine (Baltimore), 7(3).

[9] Hossam Faris, Maria Habib, Mohammad Faris, Haya Elayan, Alaa Alomari, An intelligent multimodal medical diagnosis system based on patients' medical questions and structured symptoms for telemedicine,Informatics in Medicine Unlocked, Volume 23,2021.

[10] Ahsan MM, Luna SA, Siddique Z. Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. Healthcare (Basel). 2022 Mar 15;10(3):541. doi: 10.3390/healthcare10030541. PMID: 35327018; PMCID: PMC8950225.