MACHINE LEARNING FOR DISEASE PREDICTION USING SYMPTOMS

[1]Akhil Muraleedharan, [2]Daniel K Shaji, [3]Gokul Krishna V , [4]Vinay Govindaraja Prasad, [5]S.Shankar

[1,2,3,4] Department of Computer Science and Engineering, Hindusthan College of Engineering and Technology Coimbatore, India,

20104006@hicet.ac.in, 20104014@hicet.ac.in, 20104026@hicet.ac.in, 20104063@hicet.ac.in

[5]Professor, Department of Computer Science and Engineering, Hindusthan College of Engineering and Technology Coimbatore, India, csehod@hicet.ac.in

Abstract— Using Machine learning, our project proposes a disease prediction system. For small problems, the users have to go personally to the hospital for check-up which is more time consuming. Also handling the telephonic calls for appointments is quite hectic. Such a problem can be solved by using disease prediction applications by giving proper guidance regarding healthy living. Over the past decade, the use of the specific disease prediction tools along with the concerning health has been increased due to a variety of diseases and less doctor-patient ratio. Thus, in this system, we are concentrating on providing immediate and accurate disease prediction to the users about the symptoms they enter along with the severity of disease predicted. For prediction of diseases, different machine learning algorithms are used to ensure quick and accurate predictions

Keywords— Data Processing, Machine learning Algorithms, Navie Bayes algorithms, Random Forest, Disease prediction.

I.INTRODUCTION

At present, when one suffers from a particular disease, then the person has to visit a doctor which is time consuming and costly too. Also if the user is out of reach of doctors and hospitals it may be difficult for the user as the disease can not be identified. So, if the above process can be completed using an automated program which can save time as well as money, it could be easier to the patient which can make the process easier. There are other Heart related Disease Prediction System using data mining techniques that analyzes the risk level of the patient. Disease Predictor is a web based application that predicts the disease of the user with respect to the symptoms given by the user.

The Disease Prediction system has data sets collected from different health related sites. With the help of Disease Predictor the user will be able to know the probability of the disease with the given symptoms and also a brief description of the same predicted disease is given.

As the use of the internet is growing every day, people are always curious to know different new things. People always try to refer to the internet if any problem arises. People have access to the internet more than hospitals and doctors. People do not have immediate options when they suffer with a particular disease. So, this system can be helpful to the people as they have access to the internet 24 hours.

II. LITERATURE REVIEW

I. COMPARISON OF CLASSIFICATION TECHNIQUES-SVM AND NAVIES BAYES TO PREDICT THE ARBOVIRAL DISEASE-DENGUE SHAMEEM FATHIMA et.al., has proposed in this paper In this paper we present the performance analysis of different data mining techniques to predict the Arboviral disease-Dengue. Data set used for the analysis is real time data taken from super specialty hospitals and diagnostic laboratories where the blood samples were collected for diagnostic investigations at study enrolment and again at hospital discharge.

II. FEATURE SELECTION ALGORITHMS FOR MALAYSIAN DENGUE OUTBREAKDETECTION MODEL HUSAM et.al., has proposed in this paper Dengue fever is considered as one of the most common mosquitos borne diseases worldwide. Dengue outbreak detection can be very useful in terms of practical efforts to overcome the rapid spread of the disease by providing the knowledge to predict the next outbreak occurrence.

III. EFFECTIVE ANALYSIS AND DIAGNOSIS OF LIVER DISORDER BY DATA MININGSANJAY KUMAR et.al., has proposed in this paper There are various disorders of liver that need clinical care by medical practitioner or professionals in healthcare.

IV. LIVER PATIENT CLASSIFICATION USING INTELLIGENT TECHNIQUES Anju Gulia et.al., has proposed in this paper — Classification techniques have been widely used in the medical field for accurate classification than an individual classifier. This paper presents computational intelligence techniques for Liver Patient Classification. This paper evaluates the selected classification algorithms (J-48, Multi Layer Perceptron, Support Vector Machine, Random Forest and Bayesian Network) for the classification of liver patient datasets.

V. LIVER DISEASE PREDICTION USING SVM AND NAÏVE BAYES ALGORITHMS S. Vijayarani et.al has proposed in this paper In recent years in healthcare sectors, data mining became an ease of use for disease prediction. Data mining is the process of dredge up information from the massive datasets or warehouse or other repositories.

VI. COST EFFECTIVE APPROACH ON FEATURE SELECTION USING GENETIC ALGORITHMS AND FUZZY LOGIC FOR DIABETES DIAGNOSIS. E.P.Ephzibah, et.al.,has proposed in this paper A way to enhance the performance of a model that combines genetic algorithms and fuzzy logic for feature selection and classification is proposed. Early diagnosis of any disease with less cost is preferable..

VII. DIAGNOSIS OF DIABETES USING CLASSIFICATION MINING TECHNIQUES Aiswarya Iyer et.al., has proposed in this paper Diabetes has affected over 246 million people worldwide with a majority of them being women. According to the WHO report, by 2025 this number is expected to rise to over 380 million. The disease has been named the fifth deadliest disease in the United States with no imminent cure in sight.

VIII. A HYBRID EVOLUTIONARY ALGORITHM FOR ATTRIBUTE SELECTION IN DATA MINING K.C. TAN et.al., has proposed in this paper Real life data sets are often interspersed with noise, making the subsequent data mining process difficult. The task of the classifier could be simplified by eliminating attributes that are deemed to be redundant for classification.

IX. EFFECTIVE DIAGNOSIS AND MONITORING OF HEART DISEASE Ahmed FawziOtoom et.al., has proposed in this paper Wearable sensor mobile technologies and machine learning techniques are considered as two of the key research areas in the computer science and healthcare application industries.

III. PROBLEM AND EXISTING SYSTEM

A. In the existing system, the focus is primarily on managing patient data and diagnosing diseases, particularly those of significant magnitude such as heart disease and Cancer. However, the system faces challenges due to its reliance on small datasets and pre-selected characteristics, which may not always accommodate the evolving nature of diseases and their influencing factors.

B. Patients are often required to complete lengthy questionnaires, leading to long wait times and potential inaccuracies in the results. The system's design lacks adaptability to changing disease symptoms over time, limiting its effectiveness in providing accurate diagnoses.

C. Overall, the existing system strives to address the healthcare needs of patients with specific conditions but may fall short in accommodating the dynamic nature of disease progression and evolving symptoms. Focusing on the fingertip detection and trajectory display aspects, the systemutilizes Python, OpenCV, and Convolutional Neural Networks (CNN). Python serves as the programming language, offering flexibility and ease of integration. OpenCVis utilized for precise fingertip detection, and CNNs contribute to accurate gesture recognition, allowing users to express their ideas in a virtual space with high fidelity.Comparatively, in virtual reality (VR) systems like Oculus Quest, hand controllers are commonly used for immersive interaction by tracking hand gestures.

IV. SYSTEM ARCHITECTURE

The proposed system introduces a user-friendly and efficient platform aimed at facilitating disease diagnosis and improving patient care. With a focus on simplicity and effectiveness, the system adopts a streamlined approach by accepting a maximum of five symptoms as input from users.

Featuring an intuitive and elegant user interface, the system ensures a seamless experience for both patients and healthcare professionals. Users can easily input their symptoms, minimizing the time required for data entry and enhancing accessibility to healthcare services.

Driving the diagnostic process are advanced machine learning algorithms, including Naïve Bayes, Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine. These algorithms analyze the provided symptoms to generate accurate predictions regarding potential diseases.

One of the key features of the system is its ability to offer concise descriptions of the predicted diseases. This additional information empowers users with a deeper understanding of their health conditions, enabling them to make informed decisions and take proactive measures for their well- being.

In summary, the proposed system represents a groundbreaking advancement in healthcare technology by combining user-friendly design with state-of-the-art machine learning techniques. By prioritizing simplicity, efficiency, and accuracy, the system aims to bridge the gap between patients and healthcare professionals, ultimately enhancing the quality of care and patient outcomes.

Flask: Flask framework for Python installed using the Python package manager (pip).

HTML (Frontend): HTML (Hypertext Markup Language) for creating the frontend components of the web interface.

Additional Python Libraries: Depending on the specific functionalities of the proposed system, additional Python libraries may be required for data processing, machine learning, and web development. These libraries can be installed using pip as needed.

Integrated Development Environment (IDE): An IDE such as PyCharm, Visual Studio Code, or Sublime Text for developing and testing Python and Flask applications.

Web Browser: Any modern web browser (e.g., Google Chrome, Mozilla Firefox) for accessing and testing the Flask-based web application locally.

V. ARCHITECTURE DIAGRAM

A block diagram shows the architecture of Disease Prediction

I. Block diagram:

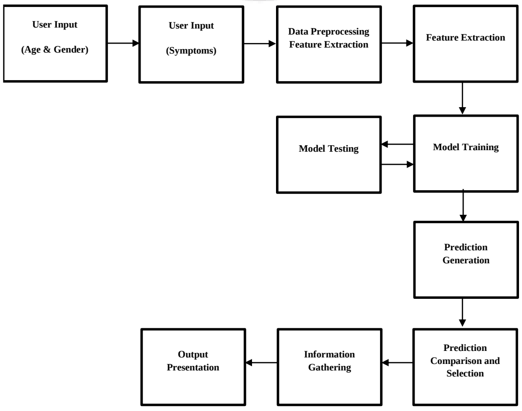


Fig. 2. Block Diagram of system

VI. IMPLEMENTATION AND DEPLOYMENT

Implementing a disease prediction system involves several key steps:

- Data Acquisition: Gather patient data from various sources, including electronic health records, medical imaging scans, and patient surveys.
- Data Pre-processing: Clean and preprocess the patient data to remove noise, standardize formats, and extract relevant features for disease prediction.
- Algorithm Selection: Choose appropriate machine learning algorithms, such as linear regression, to model the relationship between input features and disease outcomes.
- Model Training: Train the selected algorithms using labeled data to learn patterns and relationships between input features and disease probabilities.
- Model Evaluation: Evaluate the performance of the trained models using metrics such as accuracy, precision, recall, and F1-score.
- Integration: Integrate the trained models into the disease prediction system, ensuring seamless interaction with input interfaces and output generation modules.
- Continuous Improvement: Monitor the performance of the system over time and incorporate feedback from users and healthcare professionals to refine and improve disease prediction accuracy and usability.

VII. RESULTS AND DISCUSSION

Result analysis of the proposed disease prediction system involves evaluating the performance of the system based on various criteria such as accuracy, precision, recall, and F1-score. Here's a detailed description of the result analysis process:

1. Inputs:

- Patient Data: The system takes inputs in the form of patient data, which may include symptoms, medical history, demographic information, lifestyle factors, and environmental exposures.
- Feature Selection: Before applying machine learning algorithms, the system may perform feature selection to identify the most relevant attributes from the input data that contribute to disease prediction.

43

2. Algorithms Used:

- Machine Learning Models: The system employs multiple machine learning algorithms such as Support Vector Machines (SVM), Naive Bayes, Decision Trees, Random Forest, and others for disease prediction.

- Naïve Bayes: Implemented by calculating the conditional probabilities of symptoms given each disease class using the Bayes' theorem and assuming independence between features. The model estimates the likelihood of each disease given a set of observed symptoms, making it computationally efficient and effective for classification tasks in the disease prediction system
- Logistic Regression: Implemented to model the probability of occurrence of each disease class based on the input symptoms. The logistic regression model learns the relationship between the independent variables (symptoms) and the binary outcome (presence or absence of a disease) by estimating coefficients using optimization techniques such as gradient descent. It provides interpretable results and is well-suited for binary classification tasks in disease prediction.
- Decision Trees: Implemented by constructing a tree-like structure where each internal node represents a decision based on a feature, and each leaf node represents a class label (disease). The decision tree algorithm recursively partitions the dataset based on the most informative features, optimizing split criteria such as information gain or Gini impurity. Decision trees are interpretable and allow for complex decision boundaries, making them suitable for capturing nonlinear relationships between symptoms and diseases in the prediction model.

 Random Forest: Implemented as an ensemble of decision trees, where multiple trees are trained on random subsets of the dataset with replacement. Each tree in the random forest algorithm contributes to the final prediction by a voting mechanism, and the mode of the predicted classes across all trees is chosen as the final prediction. Random forests mitigate overfitting and improve prediction accuracy by introducing randomness in the training process, making them robust and effective for disease prediction in the presence of noisy or correlated features.

 Support Vector Machine (SVM): Implemented to find the hyperplane that best separates different disease classes in a high-dimensional feature space. SVM aims to maximize the margin between classes while minimizing classification errors, effectively distinguishing between different disease

patterns. SVM can handle both linear and nonlinear classification tasks by employing different kernel functions, such as linear, polynomial, or radial basis function (RBF) kernels. SVMs are versatile and suitable for disease prediction tasks with complex decision boundaries and high-dimensional feature spaces.

- Training Phase: The algorithms are trained on a labeled dataset containing historical patient data, where each instance is associated with a known disease outcome.
- Testing Phase: After training, the algorithms are tested on a separate dataset to evaluate their performance in predicting disease outcomes accurately.

3. Result Evaluation:

Accuracy: The overall accuracy of the system is calculated as the ratio of correctly predicted instances to the total number of instances tested. It provides an overall measure of how well the system performs across all classes.

- Precision and Recall: Precision measures the proportion of correctly predicted positive instances out of all instances predicted as positive, while recall measures the proportion of correctly predicted positive instances out of all actual positive instances.
- F1-Score: The F1-score is the harmonic mean of precision and recall, providing a balanced measure of a classifier's performance.

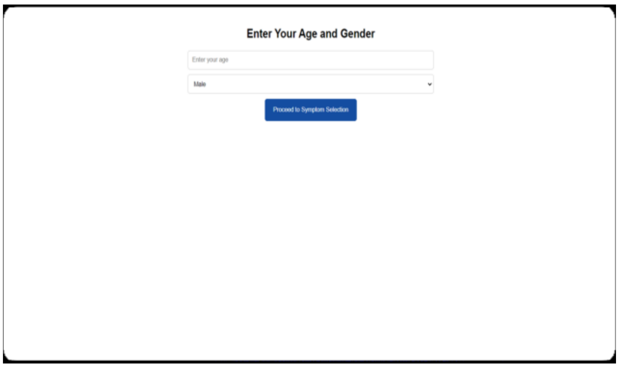
4. Output Presentation:

- Classification Labels: The system outputs the predicted disease labels for each patient, indicating the likelihood or probability of the patient having a particular disease.
- Confidence Scores: Along with classification labels, the system may also provide confidence scores associated with each prediction, indicating the level of certainty or uncertainty in the prediction.
- Visualizations: The system may generate visualizations such as confusion matrices, ROC curves, and precision-recall curves to help interpret the performance of the algorithms.

5. Interpretation and Iteration:

- Interpretation: The results are analyzed to identify patterns, trends, and areas of improvement in the system's performance. This involves understanding the strengths and weaknesses of different algorithms and feature sets.
- Iteration: Based on the result analysis, the system may undergo iterative improvements, such as refining feature selection techniques, fine-tuning algorithm parameters, or exploring ensemble methods to further enhance prediction accuracy.

Overall, result analysis in the proposed disease prediction system involves a systematic evaluation of the system's performance using various metrics and techniques, with the goal of continually improving the accuracy and effectiveness of disease prediction.

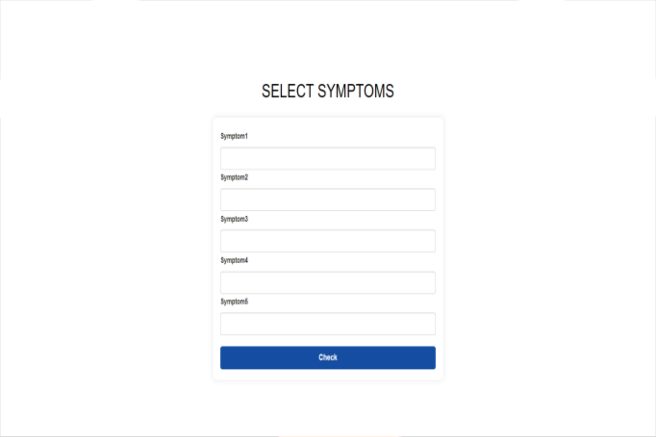


Enter Your Age and Gender

Enter your age

Male

Fig. 3. Entering age and gender



SELECT SYMPTOMS

Symptom1

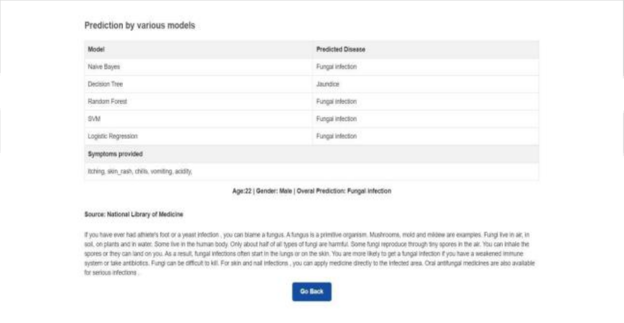
Symptom2

Symptom3

Symptom4

Symptom5

Fig. 4. Entering Symptoms for Prediction



Prediction by various models

Model	Predicted Disease
Naive Bayes	Fungal infection
Decision Tree	Jaundice
Random Forest	Fungal infection
SVM	Fungal infection
Logistic Regression	Fungal infection
Symptoms provided	
itching, skin_rash, chills, vomiting, acidity,	

Age:22 | Gender: Male | Overal Prediction: Fungal infection

Source: National Library of Medicine

If you have ever had athlete's foot or a yeast infection , you can blame a fungus. A fungus is a primitive organism. Mushrooms, mold and mildew are examples. Fungi live in air, in soil, on plants and in water. Some live in the human body. Only about half of all types of fungi are harmful. Some fungi reproduce through tiny spores in the air. You can inhale the spores or they can land on you. As a result, fungal infections often start in the lungs or on the skin. You are more likely to get a fungal infection if you have a weakened immune system or take antibiotics. Fungi can be difficult to kill. For skin and nail infections , you can apply medicine directly to the infected area. Oral antifungal medicines are also available for serious infections .

Fig. 5. Prediction using various algorithms

VIII. CONCLUSION

Overall, result analysis in the proposed disease prediction system involves a systematic evaluation of the system's performance using various metrics and techniques, with the goal of continually improving the accuracy and effectiveness of disease prediction. Virtual whiteboarding experience, Mark Air transcends traditional methodsof collaboration, offering users an interactive space where ideas can flow freely. The platform is not merely a tool; it represents a paradigm shift in how we approach virtual collaboration. It opens the door

REFERENCES

[1] Fathima, A.S. and Manimeglai, D. (2020) Predictive Analysis for the Arbovirus Dengue using SVM Classifications. International Journals of Engineering and Technology, 2, 521-527.

[2] Tarmizi, N.D.A et al., (2021) 2 feature selection algorithms for malaysian dengue outbreak detection model. Journals of Next Generation Information Technology (JNIT), 4, 96- 107.

[3] 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.

[4] Gulia, A et al., (2020) Liver Patients Classification Using Intelligent Technique. (IJCSIT) International Journal of Computer Science and Information Technology, 5, 5110-5115.

[5] Vijayarani, S. and Dhayanand, S. (2021) Liver Diseases Prediction using SVM and Naive Bayes Algorithm. International Journals of Science, Engineering and Technology Researches (IJSETR), 4, 816-820.

[6] Sarwar, A. and Sharma, V. (2020) Intelligent Naive 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] Iyer, A., Jeyalatha, S. and Sumbaly, R. (2019) Diagnosis of Diabetes Using Classification Mining Technique. International Journal of Data Mining & Knowledge Management Process (IJDKP), 5, 1-14.

[8] Tan et al., (2021) A Hybrid Evolutionary Algorithm for Attribute Selections in Data Mining. Journal of Expert Systems with Application,

[9] Vembandasamy et al., (2020) Heart Disease Detection Using Naive Bayes Algorithms. IJISET- International Journal of Innovative Science, Engineering & Technology, 2, 441-44

[10] Otoom et al., (2020) Effective Diagnosis and Monitoring of Heart Diseases. International Journal of Software Engineering and Its Application. 9, 143-156.10.1109/ICIDCA56705.2023.10099581.

[11] Aditya Tomar, “Disease Prediction System using data mining techniques”, in International Journal of Advanced Research in computer and Communication Engineering, ISO 3297, July 2016.

[12] Dr. B.Srinivasan, K.Pavya, “A study on data mining prediction techniques in healthcare sector”, in International Research Journal of Engineering and Technology (IRJET), March-2016.

[13] Megha Rathi, Vikas Pareek, “An integrated hybrid data mining approach for healthcare” , in IRACST - International Journal of Computer Science and Information Technology Security (IJCSITS), ISSN: 2249-9555 , Vol.6, No.6,Nov-Dec 2016.

[14] Feixiang Huang, Shengyong Wang, and Chien-Chung Chan, “Predicting Disease By Using Data Mining Based on Healthcare Information System” , in IEEE 2012

[15] M.A. Nishara Banu,B Gomathy, “An approach to devise an Interactive software solution for smart health prediction using data mining, in International Journal of Technical Research and Applications , eISSN, Nov-Dec 2013.

[16] Al-Aidaroos, K., Bakar, A., & Othman, Z. (2012). Medical Data Classification with Naive Bayes Approach. Information Technology Journal.

[17] Darcy A. Davis, N. V.-L. (2008). Predicting Individual Disease Risk Based On Medical Hi