JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

MULTIPLE DISEASE PREDICTION SYSTEM USING MACHINE LEARNING

[1]Dr. R Amutha, [2]Karthik M C, [3]Rohit M Sank, [4]BinduShree Y V, [5]Govardhan H

[1]Professor & Head of Department, [2]UG Student, [3]UG Student, [4]UG Student, [5]UG Student
Department of Information Science and Engineering,
AMC ENGINEERING COLLEGE, Bangalore, India

Abstract: Machine learning and artificial intelligence have become integral components of modern society, permeating various sectors such as self-driving vehicles and healthcare. The medical field, in particular, accumulates vast amounts of patient data, providing fertile ground for diverse analytical approaches. Leveraging machine learning, we've developed a Prediction System capable of detecting multiple diseases simultaneously. This project represents a significant advancement over single-disease prediction systems known for their low accuracy. Ensuring accurate predictions is crucial, as accuracy if it is low then one can pose serious risks to patients' well-being. Our system utilizes machine learning algorithms such as Random Forest Classifier and Support Vector Machine (SVM). These models are deployed using Streamlit Cloud and the Streamlit library, offering a user-friendly interface for disease prediction. The application interface offers options for three diseases: diabetes, heart disease, and Parkinson's disease. Upon selecting a disease, users are prompted to input relevant parameters for the prediction model. This initiative addresses the critical need for precise disease prediction, facilitating early detection and intervention. Furthermore, after predicting the disease, the system incorporates a pill reminder feature to alert patients about their medication schedule, enhancing adherence and overall health management.

Keywords: Machine Learning, Random Forest Classifier, Support Vector Machine (SVM), Streamlit, Diabetes, Heart, Parkinsons.

I. INTRODUCTION

In these recent years, machine learning has made big strides, especially in healthcare. Imagine if we could predict several diseases at one website without traversing to any other application using smart computer program. This could totally change how we diagnose illness and help patients get recover faster. This study looked into using a type of machine learning called Random Forest and Support Vector Machines (SVM) to predict the diseases like: Diabetes diseases, heart diseases and Parkison's diseases. To implement the analysis of multiple diseases, we'll be utilizing machine learning algorithms along with Streamlit. When users access our API, they'll need to provide the parameters of the disease along with its name. These are the serious health problem that affect a lot of people worldwide. The idea is to catch these diseases early, which can help patients. Diabetes diseases and heart disease implemented Random Forest algorithm and Parkinsons diseases implement Support vector Machine (SVM) algorithm. Random Forest is a machine learning algorithm It's like having a bunch of decision-making trees working together to solve a problem. SVM is like a super-smart tool that can sort through tons of data to find patterns. It's especially good at figuring out if someone has a disease or not based on their info. By using this machine learning algorithms accurately predicting these diseases early on, doctors can make better treatment plans and even prevent them from getting worse. Plus, it can help healthcare systems use their resources more efficiently and final model's behavior will be saved as a python pickle file.

1.1 Description

In many existing healthcare systems, you can only analyze one disease at a time. For instance, one system might focus solely on diabetes, while another handles diabetes retinopathy, and yet another predicts heart disease. This means organizations often have to deploy multiple models to analyze their patients' health reports thoroughly. But with a multiple diseases prediction system, users can analyze more than one disease on a single website. Instead of bouncing around different places, users can select the disease they're interested in, enter its parameters, and simply click submit. The system then runs the corresponding machine learning model, predicts the output, and displays it on the screen. It's like having all your disease predictions in one convenient spot.

1.2 Problem Statement

One challenge faced in developing a multiple disease prediction system for diabetes, heart and Parkinson's diseases is ensuring the accuracy and reliability of the prediction across different health care condition. Each of these diseases has unique characteristics and requires specific parameters for accurate diagnosis. Therefore, designing system that can effectively handle diverse data requirement of these disease in crucial

1.3 Proposed System

The proposed multiple disease prediction system aims to revolutionize healthcare by providing accurate predictions for diabetes, heart disease, and Parkinson's disease simultaneously. This innovative system streamlines the diagnostic process, offering users a single platform to assess multiple health conditions quickly and efficiently. and users interact with the system through a user-friendly interface designed for simplicity and ease of use. Upon accessing the system, users select the disease they wish to evaluate and input relevant parameters, such as symptoms, medical history, and diagnostic test results. With just a few clicks, the system initiates the appropriate machine learning model corresponding to the selected disease, generating accurate predictions in real-time.

II. LITERATURE SURVEY

1. The paper highlights diabetes as a significant and potentially dangerous disease due to its associated complications, such as blindness. To address the challenge of early detection, the researchers employed machine learning techniques, which offer flexibility and accuracy in predicting the presence of diabetes in patients. Their objective was to develop a system capable of accurately detecting diabetes, thereby facilitating timely intervention and treatment for affected individuals. The study focused on evaluating the performance of four primary algorithms: Decision Tree, Naïve Bayes, and Support Vector Machine (SVM). Through comparative analysis, the researchers determined the accuracy rates of these algorithms to be 85%, 77%, and 77.3%, respectively. [1]
2. In the healthcare sector, the abundance of patient data has full filled the development of advanced prediction systems leveraging machine learning techniques. Unlike many existing systems limited to single disease prediction, our newly devised Prediction System stands out by detecting multiple diseases simultaneously, thereby enhancing diagnostic efficiency and patient care. Traditional systems often suffer from lower accuracy rates, posing potential risks to patient health. In contrast, our system boasts impressive accuracy levels, achieving 89% accuracy in predicting diabetes, 83% accuracy in detecting heart diseases, and 73% accuracy in identifying liver diseases [2].
3. Machine learning models have found widespread application in disease prediction across diverse domains. Liang et al. (2019) utilized Support Vector Machine (SVM) to predict multiple diseases by analyzing electronic health records, showcasing the model's effectiveness in identifying disease patterns. Likewise, Deo (2015) employed SVM for disease prediction using clinical data, underscoring the significance of feature selection and model optimization techniques. These studies underscore the relevance and efficacy of machine learning algorithms in disease prediction [3].
4. The review of existing literature for this research project emphasized the expanding understanding of disease prediction using machine learning, with a particular emphasis on Support Vector Machine (SVM) models. It also involved comparative evaluations with alternative machine learning algorithms. This survey provided insights into the strengths and weaknesses of different approaches, aiding in the selection of appropriate methodologies for the project's objectives.[4]

5. The system addresses the significant threat posed by liver diseases, which contribute to a high number of fatalities in India and are deemed life-threatening worldwide. Early detection of liver disease is challenging but crucial for effective intervention. Leveraging automated programs employing machine learning algorithms offers a promising solution for accurate detection. The study compared the performance of Support Vector Machine (SVM), Decision Tree, and Random Forest algorithms. Precision, accuracy, and recall metrics were utilized for quantitative assessment, yielding accuracy rates of 95%, 87%, and 92%, respectively [5]

6. The paper's primary aim is to ensure accurate diagnosis and prediction of heart-related diseases, acknowledging the heart's vital role in living organisms and the potentially fatal consequences of heart-related ailments. Machine learning and artificial intelligence are leveraged to predict various natural events, including heart diseases. The study evaluates the accuracy of machine learning techniques in predicting heart disease, employing k-nearest neighbor, decision tree, linear regression, and Support Vector Machine (SVM) algorithms. Using the UCI repository dataset for training and testing, the algorithms are compared, yielding accuracy rates of 83% for SVM, 79% for decision tree, 78% for linear regression, and 87% for k-nearest neighbour [6].

7. Diabetes presents a substantial global health challenge, often leading to severe complications like heart failure, vision impairment, and kidney diseases. Patients endure multiple visits to diagnostic centers, investing time and money in obtaining reports after consultations. However, Machine Learning advancements offer a promising solution. Using sophisticated data processing techniques, predictive models can forecast an individual's likelihood of developing diabetes, enabling proactive intervention. Data mining extracts valuable insights from vast diabetes-related data. This study aims to develop a predictive system accurately assessing diabetes risk. Various classification algorithms—Decision Tree, Artificial Neural Network (ANN), Naive Bayes, and Support Vector Machine (SVM)—were utilized. Results show the Decision Tree model achieved 85% precision, while Naive Bayes and SVM reached 77% and 77.3% precision, respectively. These outcomes underscore the methods' significant accuracy in predicting diabetic risk levels. SVM particularly shines, especially with diverse or unknown data characteristics, including unstructured and semi-structured formats like text, images, and trees. This research highlights machine learning's potential in enhancing diagnostics and facilitating early diabetes management intervention. [7]

III. DESIGN

3.1 ARCHITECTURE DESIGN

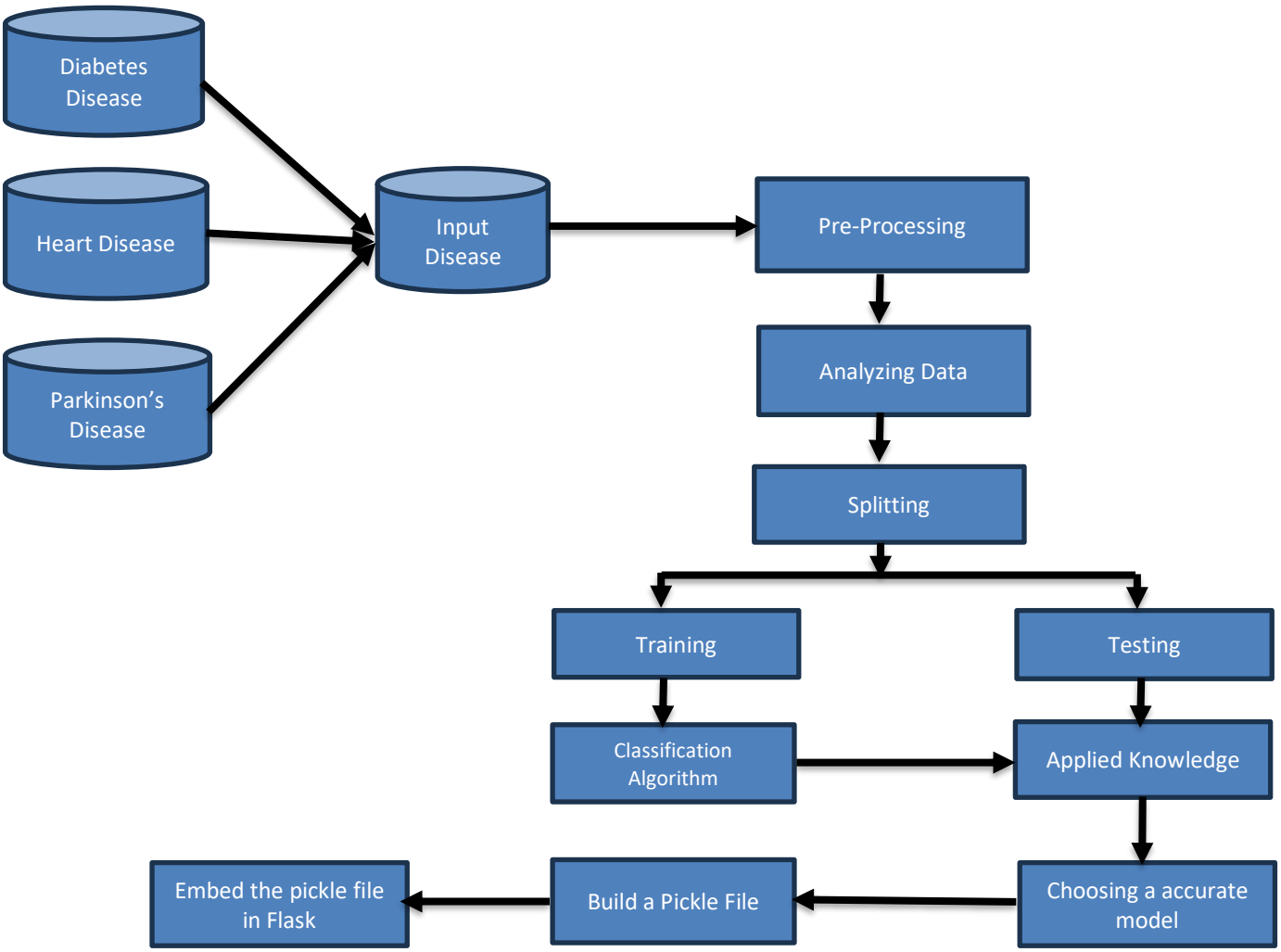

Figure 1.1 Block Diagram for Multiple Disease Prediction System

In our experiment, we focused on analyzing three diseases: heart disease, diabetes, and liver disease, which are known to be correlated with each other. To begin, we imported relevant datasets for each disease: the UCI dataset for heart disease, the PIMA dataset for diabetes, and Parkinsons's Dataset. Once the datasets were imported, we initiated the visualization process for each dataset to gain insights into the data. Following visualization, we performed pre-processing on the data, which involved identifying and handling outliers, addressing missing values, and scaling the datasets as necessary. After pre-processing, we divided the data into training and testing sets. Subsequently, we applied three different classification algorithms—Random Forest, XG Boost and Support Vector Machine (SVM)—to the training dataset. Using the testing dataset, we evaluated the performance of each algorithm by applying relevant metrics and techniques. Based on the results, we selected the algorithm that demonstrated the highest accuracy for each disease. To ensure the accessibility and usability of our models, we saved the chosen algorithms into pickle files. These pickle files were then integrated into the Django framework, enabling the models to be deployed and accessed via a webpage interface. This integration allows users to input relevant data and receive model predictions for the respective diseases on the webpage.

IV. IMPLEMENTATION

In our Health care Project, we amalgamated both structured and unstructured data to gauge disease risk. To address missing data in medical records sourced from online platforms, we employed a latent factor model. Additionally, we leveraged statistical information to evaluate prevalent diseases within specific populations and regions. To ensure our analysis was comprehensive, we collaborated with hospital specialists to identify pertinent features when handling structured data. Meanwhile, for unstructured text files, we employed the random forest algorithm to autonomously select relevant features. This approach allowed us to effectively analyze and interpret both structured and unstructured data to assess disease risk accurately.

4.1 Data Collection

For our data collection process, we sourced information from the internet to identify specific diseases. We made sure to collect real symptoms of the diseases, without inserting any dummy values. These symptoms were gathered from various reputable health-related websites. This meticulous approach ensured that our dataset accurately reflected the symptoms associated with each disease, contributing to the reliability and authenticity of our research findings.

4.2 Data Preprocessing

Prior to inputting the data into the prediction model, we execute several data cleaning and preprocessing procedures. These steps ensure the data's quality and enhance the model's performance:

- Checking for null values and utilizing the forward fill method for filling any missing values.
- Standardizing the data by adjusting it to have a mean of zero and a standard deviation of one. This step aids in comparing variables with different scales and distributions.
- Dividing the dataset into training and testing sets. This division allows us to train the model on one portion of the data and validate its performance on another, ensuring generalizability and assessing the model's ability to make accurate predictions on unseen data.

4.3 Building Model

In the process of building our model, we employ various data mining methods, with machine learning being a prominent approach. Within machine learning, techniques like random forest and Support Vector Machine (SVM) are utilized. Machine learning encompasses strategies such as grouping, clustering, and summarization, among others. Given that our project involves classification, which is a crucial aspect of data mining, we focus on categorical data classification.

4.4 Prediction

- Prediction is done by using Random Forest Algorithm and Support Vector Machine (SVM).
- Diabetes Diseases - Random Forest Algorithm
- Heart Diseases - Random Forest Algorithm
- Parkinson’s Diseases – Support Vector Machine (SVM).

4.5 Algorithm

4.5.1 Random Forest Algorithm

Random Forest operates through a two-phase process. In the first phase, it constructs a random forest by combining multiple decision trees. This process involves several steps:

Step-1: A random subset of the training data is selected. This subset typically consists of K data points randomly chosen from the entire training set.

Step-2: A decision tree is built using the selected subset of data points. Each tree is constructed independently of the others.

Step-3: The number of decision trees to be included in the random forest, denoted by N, is determined prior to construction.

Step-4: Steps 1 and 2 are repeated N times to create N decision trees. Each tree is built using a different random subset of the training data.

Once the random forest is constructed, it moves to the prediction phase:

Step-5: When presented with new data points, each decision tree in the forest makes its individual prediction. For classification tasks, this prediction could be the class label, and for regression tasks, it could be a numerical value.

Step-6: The final prediction for the new data point is determined by aggregating the predictions of all the decision trees. Typically, this is done by selecting the prediction that receives the most "votes" from the individual trees. In classification tasks, the category with the most votes is assigned to the new data point.

4.5.2 Support Vector Machine (SVM)

Regarding the Support Vector Machine (SVM) algorithm, its primary form is the SVM classifier. The SVM classifier establishes a hyperplane in an N-dimensional space to separate data points of different classes. Here are the steps involved:

Step-1: The SVM algorithm predicts classes, designating one class as 1 and the other class as -1.

Step-2: Like other machine learning algorithms, SVM transforms the business problem into a mathematical equation involving unknowns. These unknowns are determined by converting the problem into an optimization problem. The SVM classifier utilizes a loss function, specifically the hinge loss function, which is adjusted to maximize the margin.

Step-3: This loss function, or cost function, evaluates the discrepancy between predicted and actual classes. It aims to minimize error, but there's a trade-off between maximizing the margin and minimizing loss. To address this, a regularization parameter is introduced.

Step-4: Optimization involves adjusting weights through the calculation of gradients using calculus concepts like partial derivatives

Step-5: The gradients are updated using the regularization parameter when there's no misclassification, and the loss function is utilized in case of misclassification

V. Results and Discussions

In the disease prediction system, distinct machine learning algorithms are tailored to specific diseases based on their performance. For instance, the Random Forest algorithm is utilized for diabetes prediction and heart disease prediction, and Support Vector Machine (SVM) algorithm for Parkinson’s diseases due to their superior accuracies in respective domains.

Table:1.1 Accuracy level for all three Diseases.

Model	Algorithm	Accuracy
Diabetes Diseases	Random Forest Algorithm	99%
Heart Diseases	Random Forest Algorithm	98.54%
Parkinson’s Diseases	Support Vector Machine	87.17%

When a patient inputs parameters corresponding to a particular disease, the system evaluates whether the patient is likely to have that disease based on the provided information. Parameters are accompanied by prescribed value ranges, ensuring validity and relevance. Should the entered values fall outside the specified ranges, be invalid, or be left empty, the system prompts the user to correct them by displaying a warning sign. This mechanism helps ensure accurate and reliable predictions tailored to individual patient profiles and disease categories. and if the person is having disease the patient can add the timings to remind to take the pills (tablets).

1. Diabetes Disease

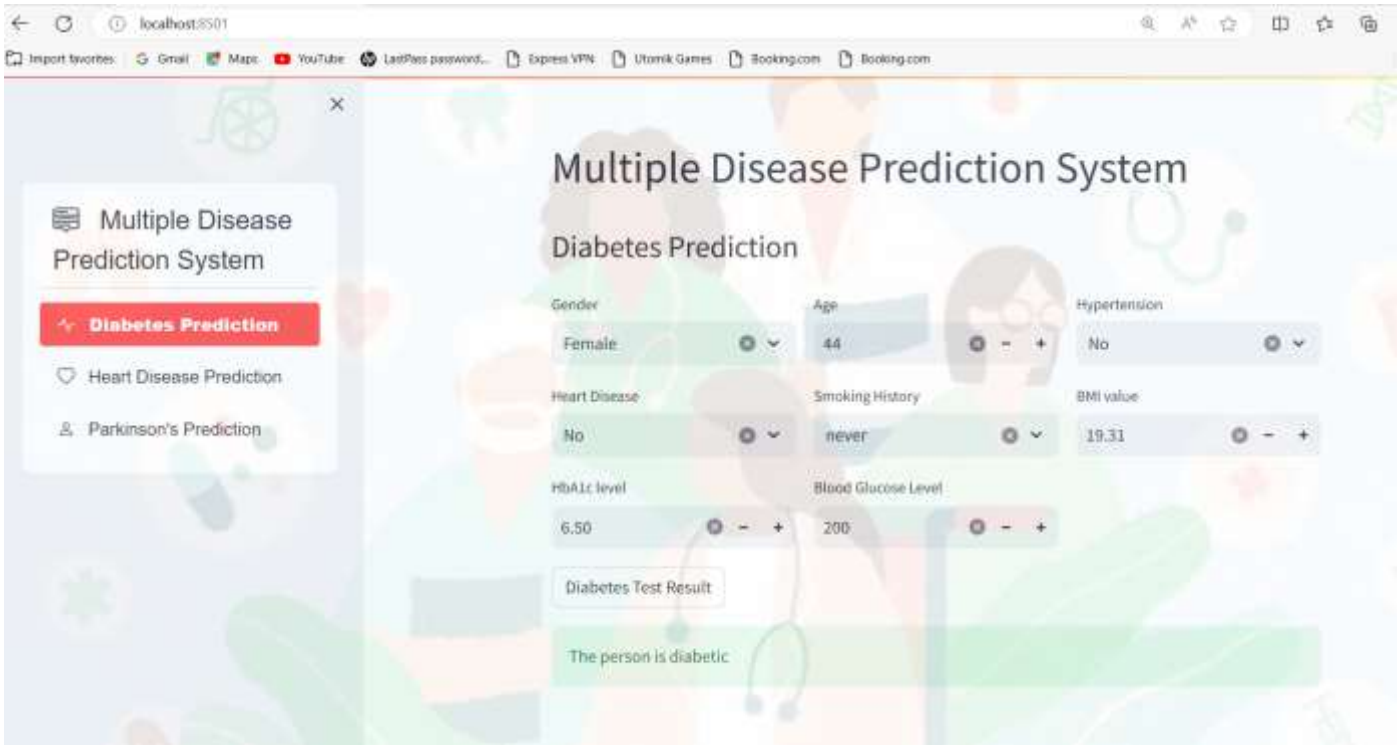

Figure 1.2: User Input Data for Diabetes Diseases

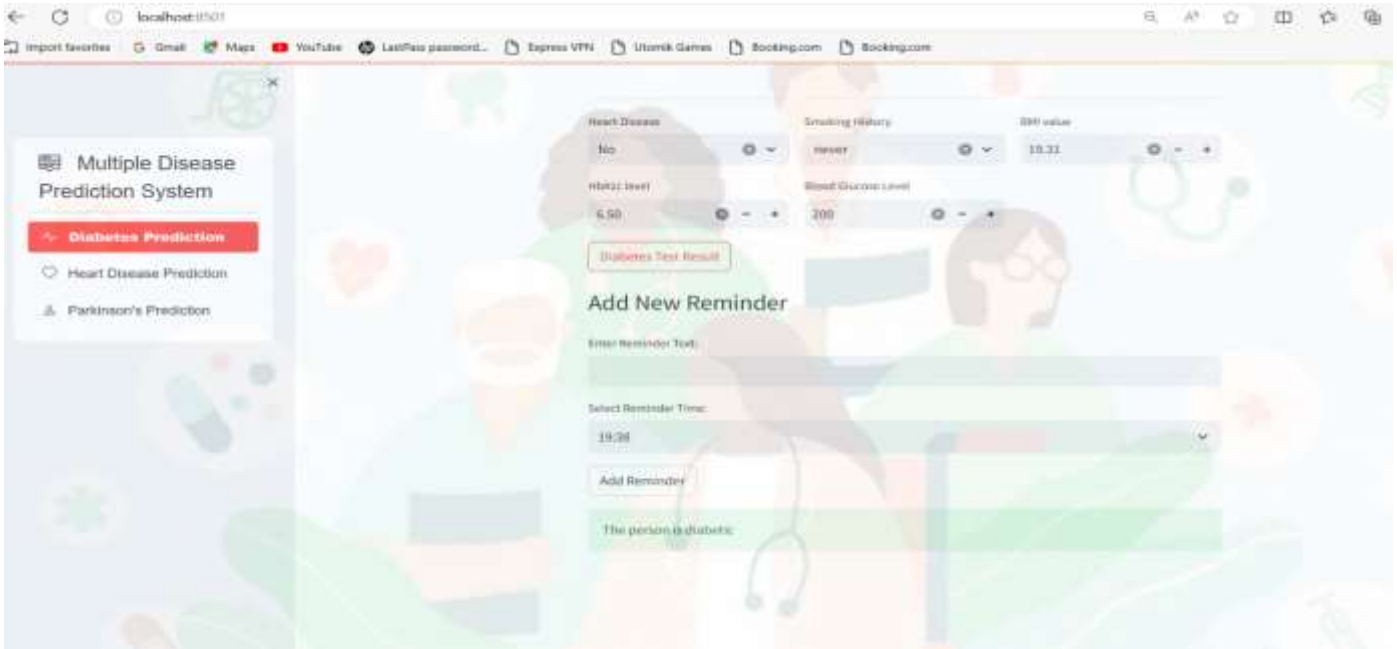

Figure 1.3: Diabetes Disease Add a Pill Reminder

2. Heart Diseases

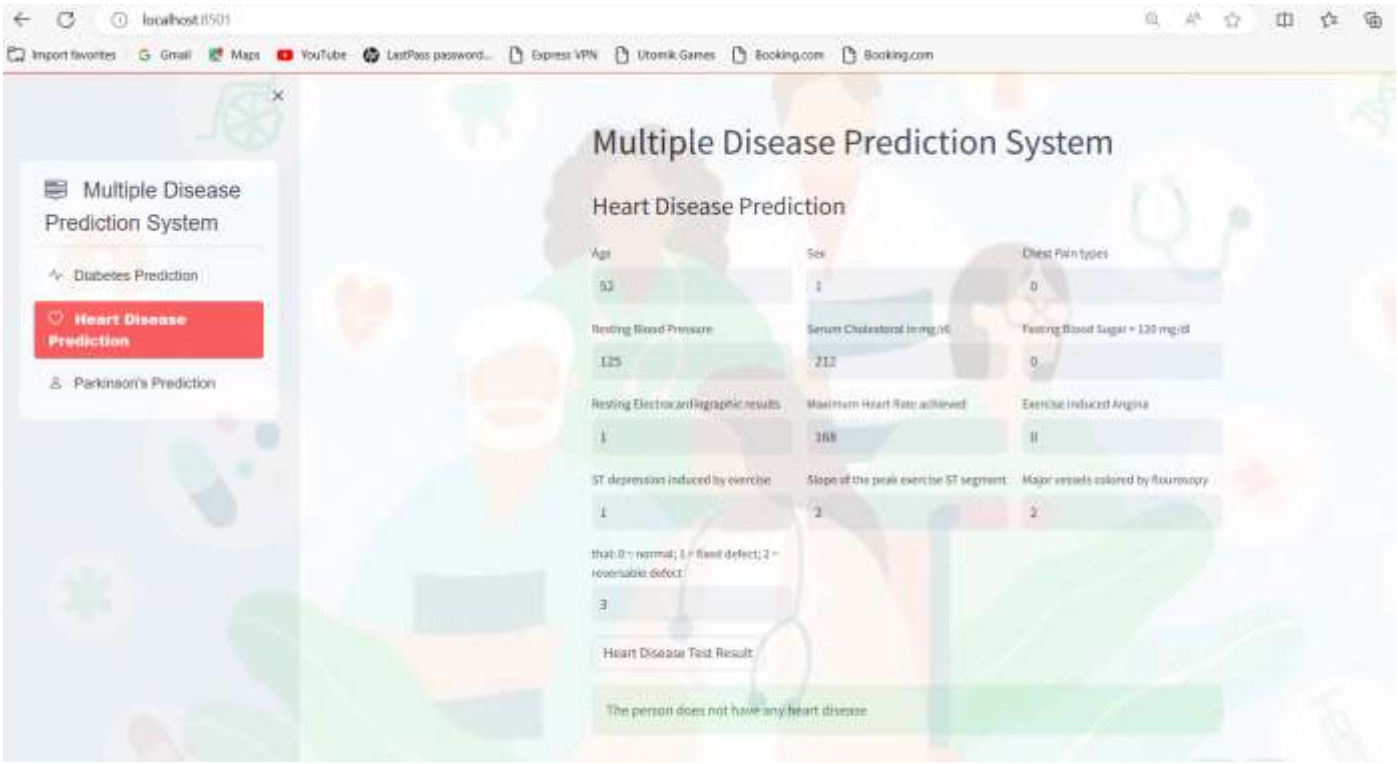

Figure 1.4: User Input Data for Heart Diseases

3. Parkinsons Disease.

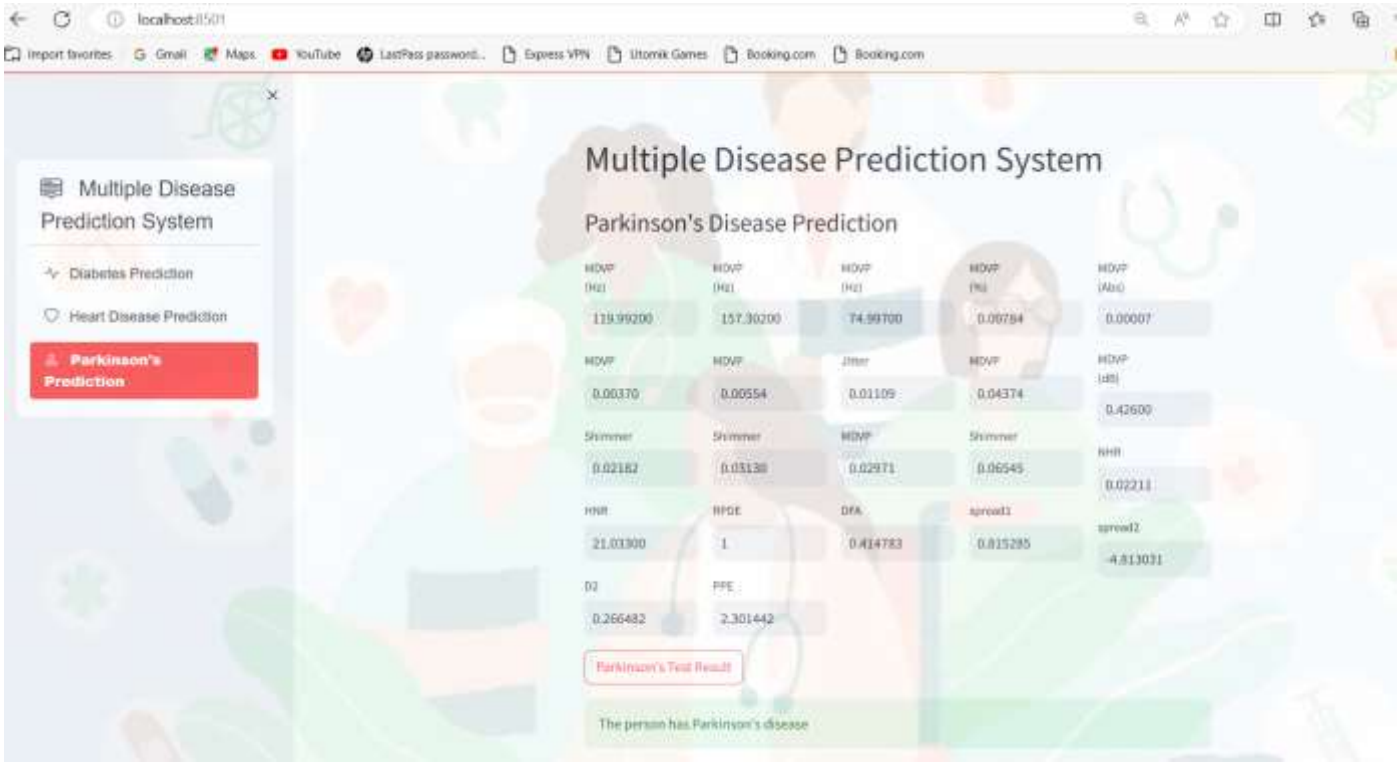

Figure 1.5: User Input Data for Parkinson's Diseases

VI. CONCLUSION

The primary goal of this project was to develop a system capable of accurately predicting multiple diseases, thereby saving users time and potentially improving their health outcomes and financial situations. By leveraging machine learning algorithms such as Random Forest and Support Vector Machine (SVM) model, the aim was to achieve high levels of accuracy in disease prediction. This system eliminates the need for users to visit various websites, streamlining the process and potentially enabling early detection of health issues. In summary, the project offers valuable contributions to healthcare accessibility and efficiency.

REFERENCES

[1] Divya Mandem, B. Prajna2 (2021) Diseases Prediction System

[2] Rudra A. Godse1, Smita S. Gunjal2, Karan A. Jagtap3, Neha S. Mahamuni4, Prof. Suchita Wankhade5 (2019) Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively.

[3] "Disease Prediction Using Machine Learning Over Big Data" Vinitha S, Sweetlin S, Vinusha H and Sajini S (2018)

[4] Ankush Singh1, Ashish Yadav2, Saloni Shah3, Prof. Renuka Nagpure4 (2022) Multiple Disease Prediction System

[5] Priyanka Sonar, Prof. K. JayaMalini," DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING APPROACHES", 2019 IEEE ,3rd International Conference on Computing Methodologies and Communication (ICCMC)

[6] Archana Singh, Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms", 2020 IEEE, International Conference on Electrical and Electronics Engineering (ICE3).

[7] A.Sivasangari, Baddigam Jaya Krishna Reddy,Annamareddy Kiran, P.Ajitha," Diagnosis of Liver Disease using Machine Learning Models" 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC).

[8] 1Mohammed Juned Shaikh, 2Soham Manjrekar, Danish Khan, 4Muzaffar Khan, 5Danish Jamadar (2022) Multiple Disease Prediction Webapp

[9] Archana Singh, Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms", 2020 IEEE, International Conference on Electrical and Electronics Engineering (ICE3)

[10] Assistant Prof. (Dr.) Jyoti Kaushik, Harshit Gupta, Lakshay Dahiya (2020) Disease Prediction System Using Machine Learning

[11] M. Nikhil Kumar, K. V. S. Koushik, K. Deepak, "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools" International Journal of Scientific Research in Computer Science, Engineering and Information Technology, IJSRCSEIT 2019.

[12] Dr. Anupam Bhatia and Raunak Sulekh, "Predictive Model for Parkinson's Disease through Naive Bayes Classification" International Journal of Computer Science & Communication vol. 9, Dec. 2017, pp. 194- 202, Sept 2017 - March 2018.

[13] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," ICT Express, vol. 7, no. 4, pp. 432–439, Dec. 2021, doi: 10.1016/j.icte.2021.02.004.

[14] M. Bayati, S. Bhaskar and A. Montanari, "Statistical analysis of a low-cost method for multiple disease prediction", Statistical Methods Med. Res., vol. 27, no. 8, pp. 2312-2328, 2018.

[15] N Chaithra and B Madhu, "Classification Models on Cardiovascular Disease Prediction using Data Mining Techniques", Journal of Cardiovascular Diseases & Diagnosis, vol. 6, pp. 1-4, 2018.