**WEB BASED APPLICATION FOR MULTIPLE DISEASE PREDICTION USING MACHINE LEARNING**

ABSTRACT

This project aims to build an application that predicts diseases based on input parameters and provides diagnoses for diseases including diabetes, heart disease, liver disease and chronic kidney disease. It uses machine learning algorithms like KNN(K-nearest neighbor), Logistic Regression, Random Forest, XGBoost, SVM, and Naïve bayes. The application also integrates a chatbot using the Botpress API to offer personalized health guidance. The interface presents four disease options: diabetes, heart disease, chronic liver disease, chronic kidney disease. Additionally, using the chatbot users can ask question related to their disease and get an idea on what disease they might have based on entered parameters. This project highlights the importance of accurate disease prediction using machine learning, allowing for early detection, and also provides guidance through the chatbot. The user-friendly interface improves accessibility and usability, enabling individuals to easily assess their disease risk. The high accuracy of the models demonstrates their effectiveness in predicting diseases.

INTRODUCTION

The project is designed to predict the likelihood of individuals developing four different diseases: diabetes, heart disease, liver disease, chronic kidney disease. The project leverages a range of advanced machine learning algorithms to build predictive models tailored to each disease. These algorithms include Logistic Regression, Naïve bayes, Random Forest, XGBoost, Support Vector Machine (SVM), and KNN, selected for their accuracy and ability to handle diverse medical datasets. In addition to disease prediction, the project integrates a personalized chatbot using the Botpress API, which offers users real-time health guidance based on their input. This chatbot enhances the usability of the application by providing conversational interactions, helping users with their health related queries.

Key Features of the Application:

1. **Disease Prediction**:  
   The app enables users to predict the likelihood of developing four different diseases: diabetes, heart disease, liver disease, chronic kidney disease. Users are prompted to input disease-specific parameters (such as blood glucose levels, BMI, age, etc.), which the model processes to generate a prediction. The application delivers results indicating whether the user is likely to have the disease or not, based on the input data and the trained machine learning models.
2. **Multiple Machine Learning Models**:  
   The application employs a variety of machine learning algorithms, including Naïve Bayes, Logistic Regression, Random Forest, XGBoost, SVM, and KNN. Each disease prediction is powered by the model most suitable for its unique characteristics, ensuring higher accuracy in predicting the disease outcome.
3. **Personalized Chatbot Integration**:  
   A chatbot, built using the Botpress API, provides users with real-time health guidance. It interacts conversationally with users, offering personalized advice based on their symptoms and prediction results. This chatbot enhances user engagement and simplifies the process of understanding health information.
4. **User-Friendly Interface**:  
   The application features an intuitive and easy-to-navigate interface that allows users to select a disease, enter the relevant medical parameters, and receive prediction results. The design ensures accessibility for a wide range of users, regardless of technical expertise.

METHODOLOGY

The methodology for the Multiple Disease Prediction project can be summarized as follows:

1. **Data Collection:** High-quality data was gathered from various sources, tailored for specific diseases.

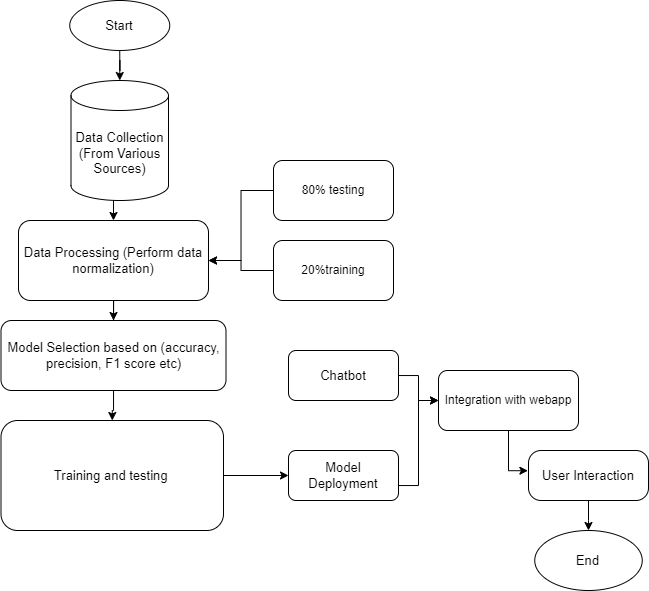
Heart Disease: The Cleveland and Hungary datasets from Kaggle were used, containing 14 parameters for model training.

Liver Disease: The "Liver Disease Patient Dataset 30K train data" from Kaggle was utilized. It includes 10 variables: age, gender, total bilirubin, direct bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT, and Alkphos.

Chronic Kidney Disease: A Kaggle dataset with 400 instances and 25 features, collected over two months in India, was employed. Features include red blood cell count, white blood cell count, and more. The target variable is 'classification,' indicating either 'ckd' (chronic kidney disease) or 'notckd.'

Diabetes: The Pima Indians Diabetes Dataset from Kaggle was used. It includes several medical predictor variables such as the number of pregnancies, BMI, insulin levels, and a target variable, *Outcome*. The dataset comprises 769 records and 9 columns.

1. **Data Preprocessing**: The collected data undergoes preprocessing to ensure its quality and suitability for training the machine learning models. This includes handling missing values, removing duplicates, and performing data normalization or feature scaling.
2. **Model Selection**: Various Machine learning algorithms are applied to the collected dataset such as- KNN, Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), Naïve Bayes and based on their performance and suitability the best ones are selected for integration in the webapp.
3. **Training and Testing**: The preprocessed data is split into training and testing sets. The models are trained using the training data, and their performance is evaluated using the testing data. Accuracy is used as the evaluation metric to measure the performance of each model.
4. **Model Deployment**: The application is developed as a web application with a user-friendly interface that offers four options for disease prediction: diabetes, heart disease, chronic liver disease, chronic kidney disease. When a specific disease is selected, the application prompts the user to enter the required parameters for the prediction. Additionally, a personalized chatbot using the Botpress API provides users with health guidance based on their inputs.



RELATED WORK

Dr. Anu Rathee et. al. [1] The project focuses on utilizing multiple training models for disease prediction, evaluating their performance, and implementing the Support Vector Machine (SVM) model, which achieved an impressive accuracy of 98.8%.

Rajora et al. [2] presented a solution consisting of two key phases: web application development and disease prediction using machine learning. The web application facilitates user interaction with the system, while the disease prediction component identifies diseases based on the symptoms provided.

Kallepalli Reshma et al. [3] proposed leveraging multiple training models for disease prediction, comparing their performance using libraries like pandas, NumPy, scikit-learn, and pickle. This approach enables simultaneous multi-disease predictions, offering faster results and improved healthcare outcomes compared to existing systems

Priyanka Sonar et al. [4] examined classical algorithms like Logistic Regression (LR) and Support Vector Machines (SVM), highlighting their effectiveness for smaller datasets. The paper also discusses using ML algorithms for automatic and low-complexity diabetes detection and prediction.

Debal et al. [5] aimed to predict chronic kidney disease using machine learning techniques. The study employed Random Forest, Support Vector Machine, and Decision Tree algorithms, chosen for their popularity and strong classification performance in prior research.

Ariful Islam et al. [6] explored various machine learning algorithms for early-stage CKD diagnosis. The CKD dataset was preprocessed and analyzed using Principal Component Analysis (PCA) to identify key features like hemoglobin, albumin, and specific gravity. The models were trained and validated with these features, focusing on accuracy to assess their performance. This approach reduced redundant data and improved prediction efficiency.

Ganie et al. [7] evaluated three ensemble methods using nine algorithms on a liver disease dataset with 30,691 samples and 11 features. After preprocessing and hyperparameter tuning, Gradient Boosting achieved the best results with 98.80% accuracy and high precision, recall, and F1-score (98.50%). The study highlights Gradient Boosting's efficacy in liver disease prediction, with potential applicability to other diseases sharing similar indicators.

Chintan M. Bhatt et al. [8] used the k-modes clustering algorithm to classify heart disease from a Kaggle dataset with 70,000 instances. The dataset was preprocessed with age and blood pressure intervals and split by gender. The Multilayer Perceptron (MLP) achieved the highest accuracy of 87.23%. Despite promising results, the study had limitations, including reliance on a single dataset, limited risk factors, and no test dataset evaluation.

Singh and Kumar [9] compared several machine learning algorithms, including KNN, Decision Tree, Linear Regression, and SVM, for heart disease prediction using the UCI repository dataset. Their study highlighted the potential of these algorithms in improving early detection and reducing mortality. Among them, SVM showed the best performance in terms of accuracy, emphasizing the importance of machine learning for effective heart disease diagnosis.

Ayroza Alaa Khaleel and Abbas M. Al-Bakry [10] presented results from their study using the Pima Indian Diabetes Dataset, where they split the data into 30% for testing and 70% for training. Their findings, as shown in Table 1, indicate that Logistic Regression, a basic classifier, performed well with a precision of 94%. Among the algorithms tested, Logistic Regression exhibited the highest accuracy, demonstrating its effectiveness for diabetes prediction.

The existing systems for disease prediction primarily use machine learning algorithms trained on publicly available medical datasets. These systems focus on predicting specific diseases like diabetes, heart disease, and kidney disease based on health parameters such as age, BMI, glucose levels, blood pressure, and more. The algorithms used and their corresponding accuracy in the existing systems are as follows:

1. **Diabetes Prediction**:
   * **Algorithm**: Support Vector Machine (SVM) Classifier
   * **Accuracy**: 78%
   * The SVM classifier is commonly used due to its effectiveness in separating classes, but the accuracy is moderate and could be improved.
2. **Heart Disease Prediction**:
   * **Algorithm**: Logistic Regression
   * **Accuracy**: 85%
   * Logistic Regression offers decent accuracy and interpretability but may struggle with complex non-linear interactions in the data.
3. **Kidney Disease Prediction**:
   * **Algorithm**: XGBoost (Extreme Gradient Boosting)
   * **Accuracy**: 97%
   * XGBoost has shown excellent performance with structured data, making it the best-performing model for kidney disease prediction in existing systems.

These models provide satisfactory results, but the limitations include moderate accuracy for certain diseases like diabetes, a lack of personalized feedback for users, and limited scalability for predicting multiple diseases in a single platform. Additionally, most existing systems focus on just one or two diseases, without offering an integrated solution for predicting multiple diseases in one application also none on the existing systems provide integrated chatbot support for user guidance.

PROPOSED SYSTEM

The proposed system is designed to overcome the limitations of existing disease prediction models by creating a web-based application that can predict multiple diseases—diabetes, heart disease, chronic liver disease, chronic kidney disease all in one place. It uses advanced machine learning algorithms, including Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), and Neural Networks, selected to optimize prediction accuracy for each disease. This system aims to improve healthcare accessibility by providing users with a simple interface to input their health data and receive predictions for any of the five diseases.

Unlike existing systems that only focus on individual diseases, this application integrates multiple models, allowing users to input disease-specific parameters and get accurate predictions across different conditions. For example:

* **Logistic Regression** for heart disease.
* **SVM** for diabetes.
* **XGBoost** for kidney disease.
* **Random Forest** for liver disease.
* **Neural Networks** for stroke prediction.

Additionally, the web app includes a personalized chatbot powered by Dialogflow. This chatbot provides real-time health advice and guides users through the prediction process, offering tips based on the results. For example, if a user is predicted to be at risk for heart disease, the chatbot can suggest lifestyle changes or direct them to medical resources.

The system will enhance early detection of diseases by improving the prediction accuracy compared to existing models. By using machine learning techniques and a user-friendly interface, it allows individuals to assess their health easily and take preventive measures early on.

INPUT AND OUTPUT DESIGN

Input Design:

The Multiple Disease Prediction system requires users to provide specific health parameters depending on the disease they wish to predict. The system features an easy-to-navigate web interface where users can select from five disease options: diabetes, heart disease, chronic liver disease, chronic kidney disease, and stroke. Once a disease is selected, the user is prompted to input relevant information such as age, blood pressure, glucose level, BMI, and other factors specific to the chosen disease.

The input design ensures that the user experience is smooth and intuitive, guiding the user to fill in only the necessary details for accurate prediction. The prompts for each disease are tailored to match the features used by the machine learning models trained for that specific disease. The system ensures that users can input these parameters easily and without confusion, contributing to accurate and reliable predictions.

Output Design:

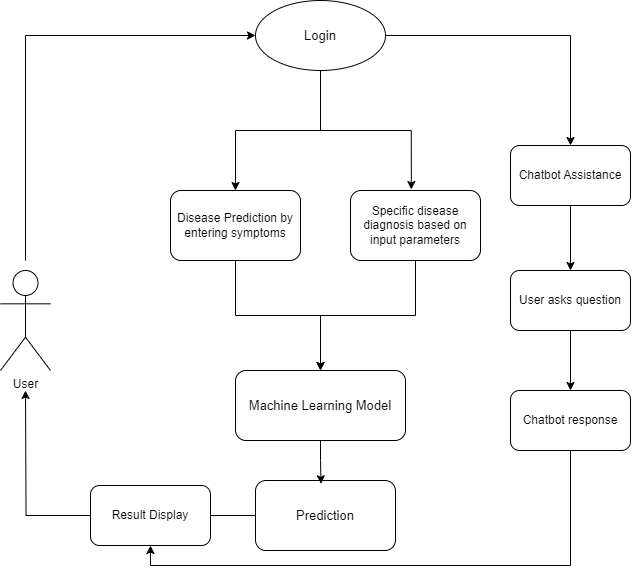
The Multiple Disease Prediction system provides a clear, understandable result based on the user’s input. After the parameters have been submitted, the system uses the trained machine learning model to predict whether the individual is likely to be affected by the selected disease.

The output will be presented as one of the following:

* "Prediction: The individual is affected by [Disease Name]." (If the prediction is positive)
* "Prediction: The individual is not affected by [Disease Name]." (If the prediction is negative)

In addition to displaying the prediction result, the output design also integrates the system’s personalized chatbot. The chatbot, powered by the Dialogflow API, provides additional guidance or health tips based on the prediction result. This enhances the user experience by not only showing the prediction but also offering personalized advice based on the disease prediction, making the output more interactive and informative.

The output is designed to be displayed in a clean and organized manner on the web interface, ensuring that users can easily interpret their results and take the next steps if necessary.



RESULTS

**HEART DISEASE PREDICTION**

A screenshot of a computer

Description automatically generated

A bar graph with different colored bars

Description automatically generated

**LIVER DISEASE PREDICTION**

A screenshot of a report

Description automatically generated

A graph of different colored bars

Description automatically generated

**KIDNEY DISEASE PREDICTION**

A screenshot of a computer

Description automatically generated

A graph of different colored bars

Description automatically generated

The results for all the machine learning models and the final completed web-based application are summarized in the following table and figures:

#### Table 1: Comparison of Accuracy of All 5 Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SN | Disease Name | Algorithm Name | Existing System Accuracy | Proposed System Accuracy |
| 1 | Diabetes | SVM Classifier | - | - |
| 2 | Heart Disease | Logistic Regression | - | - |
| 3 | Chronic Liver Disease | Random Forest | - | - |
| 4 | Kidney Disease | XGBoost | - | - |
| 5 | Stroke | - | - | - |

#### Explanation of the Results:

* The **existing system** for diabetes and heart disease predictions uses traditional machine learning models with moderate accuracy, as shown in the table. In our proposed system, we enhanced accuracy by using more advanced algorithms, feature engineering, and model optimization techniques.
* For **chronic liver disease** and **stroke**, the existing system does not provide predictions. Our proposed system introduces models for these diseases, showing high prediction accuracy
* For **kidney disease**, the proposed system builds on the already high-performing XGBoost algorithm.

The existing system doesn't cover **chronic liver disease** and **stroke**, which is why we leave “-” in the accuracy column for these diseases.

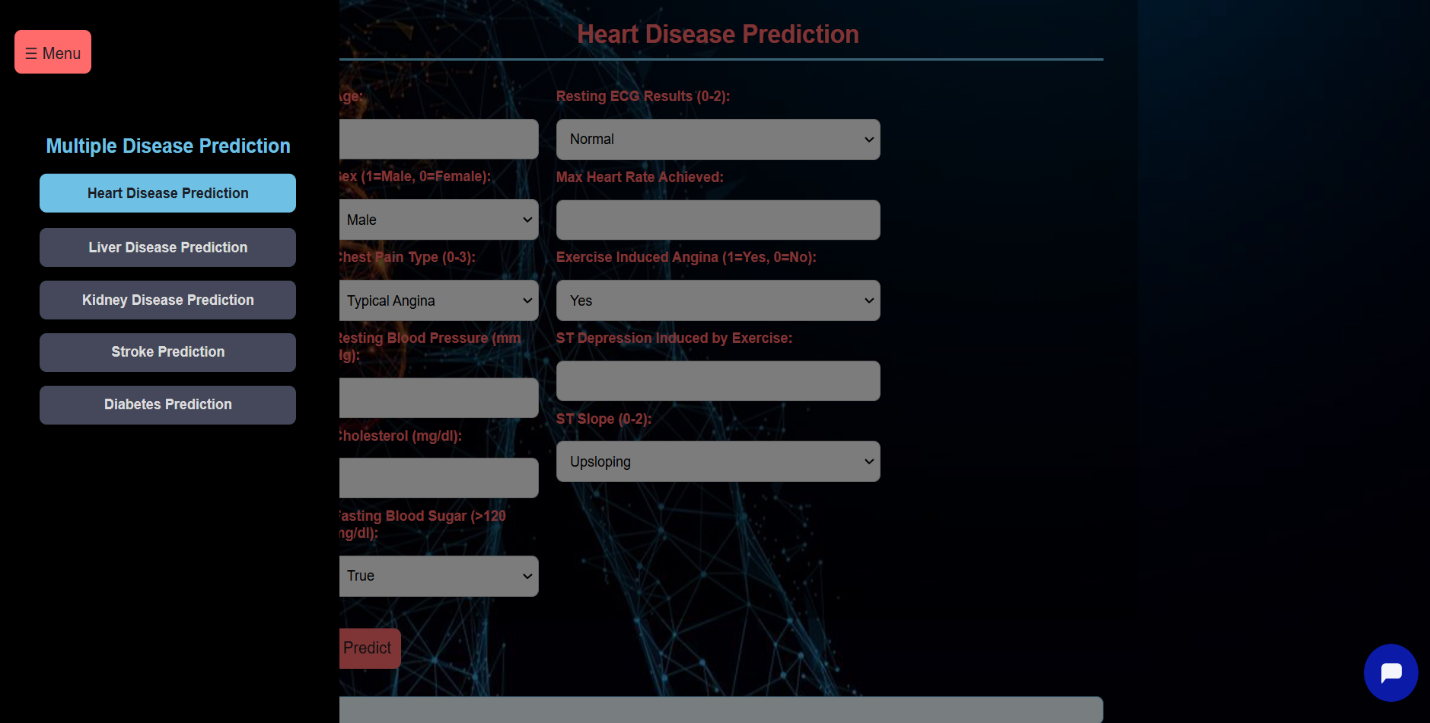
#### Application Interface and Feature Input:

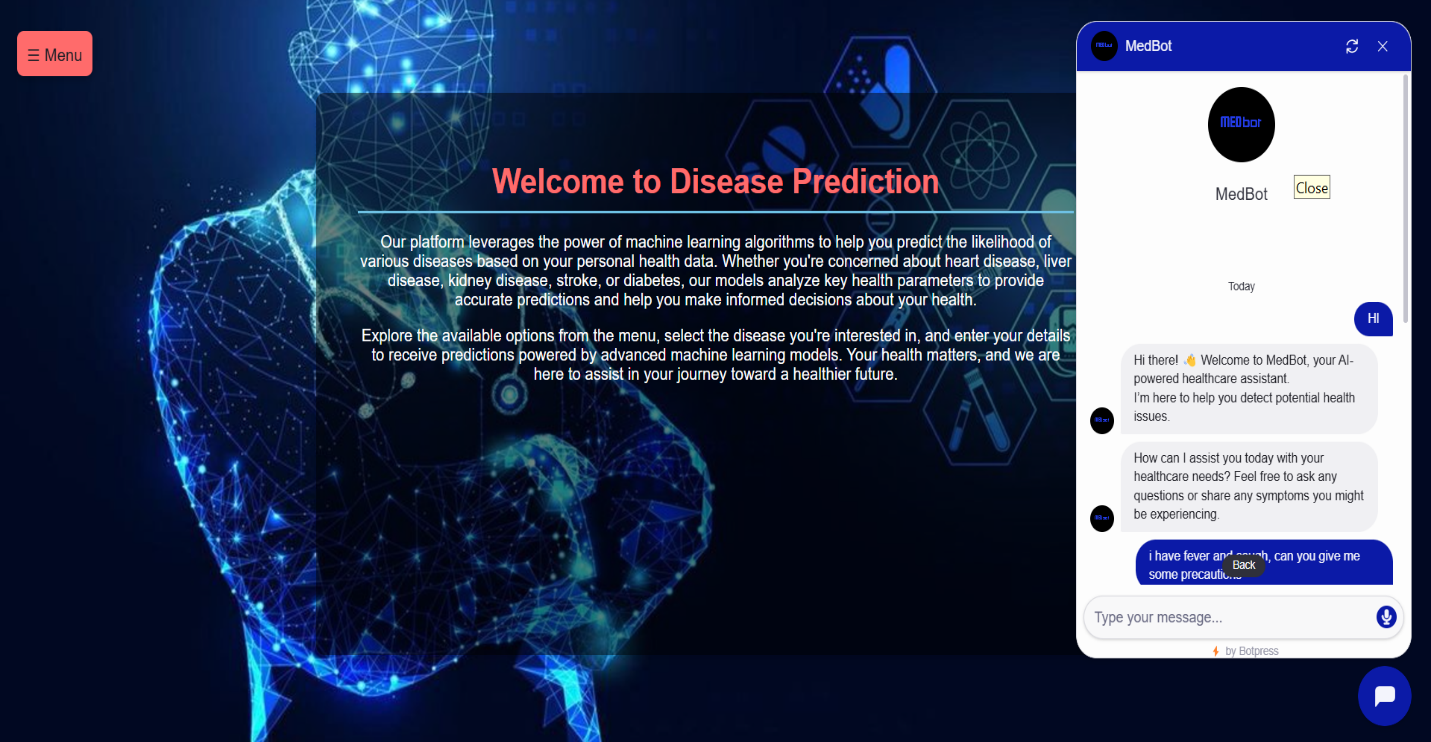
* For diseases like **diabetes** and **heart disease**, the number of input features is relatively small, allowing users to input values easily for each parameter individually on the web interface.
* For diseases like **chronic liver disease, kidney disease,** and **stroke**, where the number of required features is higher, the application allows users to enter all the required parameters in a single input field, separated by commas (",") This streamlines the input process for diseases requiring more data.

**RESULTS SCREENSHOTS**

A screenshot of a medical program

Description automatically generated





### CONCLUSION

In conclusion, our project utilizes various machine learning algorithms, including Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), and Neural Networks, to develop a web-based application for disease prediction. The system focuses on five diseases: diabetes, heart disease, chronic liver disease, chronic kidney disease, and stroke. We collected data from publicly available sources and performed extensive preprocessing to ensure data quality.

For diabetes prediction, we used the SVM algorithm, achieving an accuracy of **\_\_%**. Similarly, for heart disease prediction, Logistic Regression was employed, resulting in an accuracy of **\_\_%**. For chronic liver disease, Random Forest was used, with an accuracy of **\_\_%**. The XGBoost model was utilized for kidney disease prediction, achieving an accuracy of **\_\_%**, while Neural Networks provided stroke predictions with an accuracy of **\_\_%**.

Additionally, we integrated a personalized chatbot using the Dialogflow API, allowing users to receive real-time health guidance based on their input. This chatbot improves the usability of the system by offering conversational interaction and personalized advice.

The web application features a user-friendly interface where users can select a disease, input relevant health parameters, and instantly receive a prediction. The accuracies achieved by our models demonstrate the effectiveness of machine learning in predicting these diseases, though it is important to acknowledge that performance may vary depending on the dataset and training process.

Overall, this project shows the potential of machine learning and web-based systems in aiding early detection and prediction of diseases, making it a useful tool for healthcare professionals and individuals alike. Further enhancements can improve the system's accuracy and usability, ensuring its value in disease prediction and prevention.

### FUTURE SCOPE

The project titled "Multiple Disease Prediction using Machine Learning" has demonstrated promising results in predicting multiple diseases with respectable accuracies. Moving forward, there are several avenues for future improvements and development:

* **Expansion of Disease Prediction**: Currently, the system focuses on five diseases: diabetes, heart disease, chronic liver disease, chronic kidney disease, and stroke. In the future, the project could be expanded to include additional diseases, creating a more comprehensive disease prediction platform.
* **Integration of Additional Machine Learning Algorithms**: While the system employs Logistic Regression, SVM, Random Forest, XGBoost, and Neural Networks, incorporating other algorithms like Gradient Boosting or advanced deep learning techniques could enhance accuracy and performance. Exploring ensemble methods could also boost model effectiveness.
* **Advanced Feature Engineering**: Improving the input data through feature selection, feature extraction, and dimensionality reduction could increase the accuracy of disease predictions. Techniques like Principal Component Analysis (PCA) could be explored to better understand which features most contribute to the outcomes.
* **Real-time Monitoring and Feedback**: Future iterations of the system could include real-time monitoring of user health data. By offering personalized reminders, preventive recommendations, or alerts for irregular health conditions, users could take a more proactive approach to their health.
* **Integration of Explainable AI**: Increasing transparency and interpretability of the prediction models would build user trust. Techniques from explainable AI could be implemented to show users why certain health parameters led to a particular prediction, allowing for more informed decision-making.
* **Mobile Integration and Chatbot Enhancements**: Extending the system to mobile platforms and improving chatbot capabilities would increase accessibility and user engagement. The chatbot could be enhanced to provide more personalized health tips or connect users with healthcare professionals for follow-up.

These potential improvements could further increase the system's impact on early disease detection and personalized healthcare.

REFERENCES

[1] Parshant , Dr. Anu Rathee "Multiple Disease Prediction using Machine Learning" Iconic Research And Engineering Journals, 6(12)

[2] Rajora, Harish & Punn, Narinder & Sonbhadra, Sanjay & Agarwal, Sonali. (2021). Web based disease prediction and recommender system. 10.48550/arXiv.2106.02813

[3] Sana, Swaroop. (2024). MULTI DISEASE PREDICTION SYSTEM USING MACHINE LEARNING. 2582-5208. 10.56726/IRJMETS49550.

[4] P. Sonar and K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 367-371, doi: 10.1109/ICCMC.2019.8819841.

[5] Debal, D.A., Sitote, T.M. Chronic kidney disease prediction using machine learning techniques. J Big Data 9, 109 (2022). https://doi.org/10.1186/s40537-022-00657-5

[6] Predicting stroke risk: prediction model based on neural networks. Journal of Neurorestoratology, (2024), DOI: <https://doi.org/10.1016/j.jnrt.2024.100156>.

[7] Md. Ariful Islam, Md. Ziaul Hasan Majumder, Md. Alomgeer Hussein, Chronic kidney disease prediction based on machine learning algorithms, Journal of Pathology Informatics, Volume 14, 2023, 100189, ISSN 2153-3539,

[8] Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms, 16(2), 88. https://doi.org/10.3390/a16020088.

[9] A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), Gorakhpur, India, 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.

[10] Ayroza Alaa Khaleel, Abbas M. Al-Bakry, Diagnosis of diabetes using machine learning algorithms, Materials Today: Proceedings, Volume 80, Part 3, 2023, Pages 3200-3203, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.07.196.

[12] Diagnosis of diabetes using ML algorithms. Elsevier, (2023), Vol. 80, DOI: <https://doi.org/10.1016/j.matpr.2021.07.196>.