International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Multiple Disease Prediction Using Machine Learning and Stream- LIT

Disha Sahu, Diya Roy, Mallapuram Sneha Rao, Manisha Dalai

(B-Tech Students, Department of Computer Science Engineering, Shri Shankaracharya Technical, Campus, Bhilai Chhattisgarh)

Guide: Abhishek Kumar Dewangan

(Professor, Department of Computer Science Engineering, Shri Shankaracharya Technical Campus, Bhilai Chhattisgarh)

ABSTRACT

The Aim of this project is to develop a machine learning-based web application capable of predicting multiple diseases, specifically diabetes, heart disease, and Parkinson's disease. It utilizes classification algorithms such as Support Vector Machine (SVM) and Logistic Regression to make predictions based on user-input health data. The system is deployed using the Streamlit framework and Streamlit Cloud, offering a simple and interactive user interface. Users can select from the available disease options, enter the required medical parameters, and receive immediate prediction feedback on whether they are at risk. By incorporating reliable ML models and offering an accessible web interface, this application helps promote early diagnosis and supports timely medical intervention. The effective performance of the models, indicated by their high accuracy levels, showcases the potential of machine learning in enhancing digital healthcare solutions.

Keywords: Disease Prediction, Machine Learning, Streamlit, SVM, Logistic Regression, Early Diagnosis, Health Monitoring.

I. INTRODUCTION

This project, titled "Multiple Disease Prediction using Machine Learning and Streamlit," aims to build an intelligent system that can predict several common health conditions, including diabetes, heart disease, and Parkinson's disease. To achieve this, the system uses machine learning techniques—specifically Support Vector Machine (SVM) for predicting diabetes and Parkinson's, and Logistic Regression for identifying the likelihood of heart disease.

The datasets used to train these models were sourced from Kaggle and carefully pre-processed to ensure the data was clean and suitable for model training. Each health condition is associated with a particular algorithm based on its effectiveness with the respective dataset. Once trained, the models are integrated into a web application built using Streamlit, which offers a clean and interactive user interface.

Users can choose from a list of diseases and enter the relevant medical details required for prediction. After submission, the system instantly evaluates the inputs using the trained model and delivers a prediction indicating whether the individual may be at risk.

The application is hosted on Streamlit Cloud, which allows it to be accessed from any browser without additional installations. The combination of machine learning and a user-friendly web platform makes this tool effective for early diagnosis and risk assessment, helping users take proactive steps toward their health. interface allows users to input disease-specific parameters and obtain prediction results, facilitating early detection and proactive healthcare management.

II. METHODOLOGY

The approach followed for developing the Multiple Disease Prediction system is outlined in the following steps:

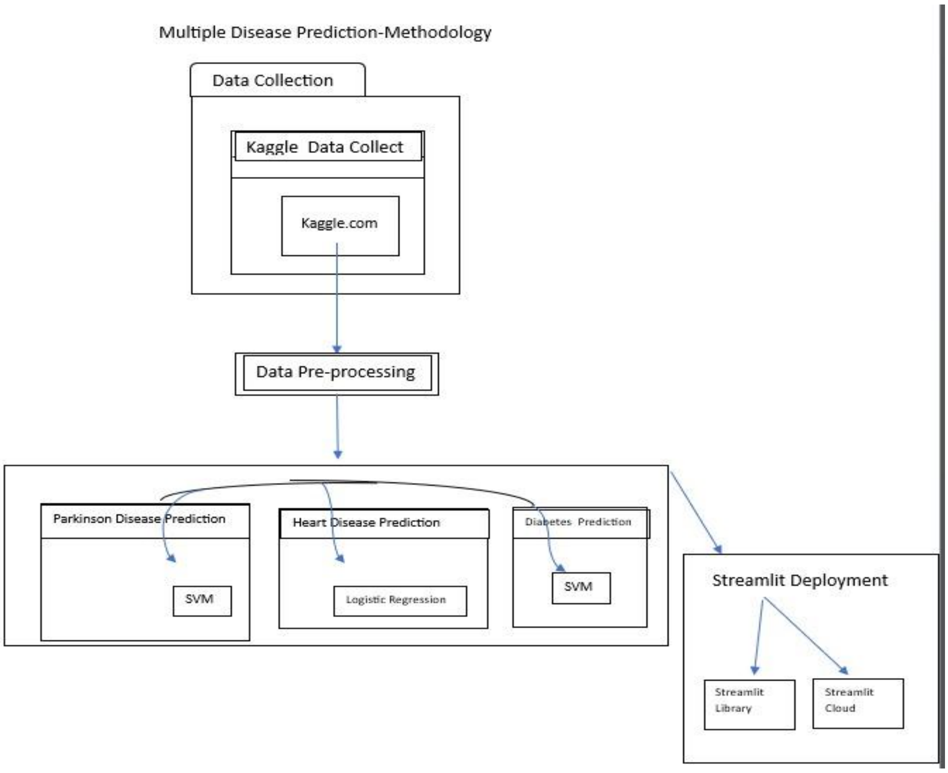
1. Data Collection: To begin with, datasets for diabetes, heart disease, and Parkinson's disease were sourced from Kaggle, a widely-used platform that provides open access to high-quality datasets for data science projects.

2. Data Preprocessing: Before feeding the data into machine learning models, several preprocessing steps were carried out to ensure the data was clean and consistent. This included handling any missing entries, removing duplicate records, and applying normalization or feature scaling techniques to bring all features to a comparable scale.

3. Model Selection: Specific machine learning algorithms were chosen for each disease based on their historical performance and compatibility with the respective datasets. Support Vector Machine (SVM) was selected for predicting diabetes and Parkinson's disease, while Logistic Regression was used for heart disease due to its efficiency and interpretability.

4. Model Training and Evaluation: The cleaned datasets were divided into training and testing subsets. The models were trained using the training data, and their effectiveness was evaluated using the test data. Accuracy was the primary metric used to assess how well each model performed.

5. Model Deployment: The final models were integrated into a web application built using Streamlit, a Python library designed for creating interactive web interfaces. The app was deployed using Streamlit Cloud, making it accessible via any browser. Users can select one of the three diseases, input relevant health information, and receive immediate predictions.



III. PROBLEM STATEMENT

The goal of this project is to create a machine learning application that uses Support Vector Machine (SVM) and Logistic Regression to predict several diseases, including diabetes, heart disease, and Parkinson's disease. The application will allow users to input specific health data and receive a prediction regarding whether they may be at risk for these conditions based on models trained with historical data. By utilizing these machine learning algorithms, the project aims to assist in early disease detection, which can lead to better healthcare outcomes. Additionally, the user-friendly interface of the application will make the prediction process accessible, helping individuals understand their health status more easily.

IV. EXISTING SYSTEM

This project focuses on developing an intelligent and user-friendly system that can predict the possibility of individuals being affected by multiple diseases, specifically diabetes, heart disease, and Parkinson's disease, using machine learning techniques. The system integrates a variety of classification algorithms, including Logistic Regression, Naive Bayes, Decision Trees, Random Forest, and Support Vector Machine (SVM), each of which approaches prediction differently, allowing the system to learn from the data in diverse ways. To make the platform accessible and easy to navigate, the entire model is built using the Streamlit framework and hosted on Streamlit Cloud, enabling users to interact with the system through a simple and efficient web interface. The process begins with the collection of relevant medical datasets from authentic and publicly available sources. These datasets undergo careful preprocessing steps such as handling missing values, normalization, and feature selection to ensure that the models receive clean and meaningful input. Once the data is ready, the system trains each machine learning model to identify patterns that can indicate the presence or absence of a particular disease. Among the models tested, the Support Vector Machine demonstrated noteworthy results, achieving an accuracy of 76% in predicting diabetes and 71% in identifying Parkinson's disease. This performance suggests that the SVM model is capable of effectively distinguishing between healthy individuals and those at risk, making it a valuable tool for early detection. Other models, such as Decision Trees and Random Forest, also play a significant role by offering alternative approaches to classification, which enhances the overall robustness and reliability of the system. By leveraging the strengths of multiple algorithms, the platform improves its chances of providing accurate and consistent results across different diseases. Overall, this project illustrates the practical application of machine learning in the healthcare domain, showing how predictive analytics can assist in early diagnosis, potentially reduce the burden on healthcare systems, and empower users to take proactive steps toward managing their health. With further improvements—such as expanding the dataset, fine-tuning the models, and incorporating real-time data—the system has the potential to evolve into a more powerful and widely adopted medical support tool.

V. PROPOSED SYSTEM

The proposed system introduces several enhancements to improve the accuracy and usability of disease prediction. It incorporates preprocessing techniques such as data standardization to ensure consistency in feature scales, and label encoding to convert categorical values into numerical format. Additionally, dimensionality reduction techniques are applied to simplify the dataset while retaining critical information, helping the models perform more efficiently.

Unlike the existing system, which relies on a Flask-based API, the new system is built using the Streamlit library and hosted on Streamlit Cloud, with the codebase managed on GitHub. This change simplifies deployment and makes the application more accessible and visually intuitive for end users.

The platform is designed to predict the likelihood of three common diseases: diabetes, heart disease, and Parkinson's disease. For this, we employ machine learning algorithms tailored to each condition. The data, sourced from Kaggle, is pre-cleaned and prepared to train these models. Support Vector Machine (SVM) is used for both diabetes and Parkinson's prediction due to its effectiveness with those datasets, while Logistic Regression is applied to predict heart disease outcomes

Performance-wise, the system shows strong results. The SVM model for diabetes prediction achieved an accuracy of 78%, indicating reliable detection of risk. For Parkinson's disease, the same algorithm reached an accuracy of 87%, showcasing its strength in differentiating between affected and unaffected individuals. The heart disease model, built using Logistic Regression, scored 85% accuracy, further supporting its practical utility in early intervention. In some cases, model performance reached as high as 97% during evaluation.

The user interface provides a simple dropdown menu to choose one of the three diseases. Upon selection, the system requests the user to input specific health parameters. Once submitted, the application processes the inputs through the relevant model and returns the prediction. This interactive and straightforward design allows users to quickly assess potential health risks and encourages timely medical consultation if needed.

VI. INPUT AND OUTPUT DESIGN

Input Design:

The system is designed to collect user input in the form of specific health-related parameters that are essential for predicting each disease. When a user selects a disease from the available list—such as diabetes, heart disease, or Parkinson's disease—the application automatically requests the relevant medical information needed for that condition. The input fields are organized in a simple and intuitive layout to make the process easy for users, even those with minimal technical background. The goal is to ensure that users can quickly and accurately enter the required data, improving the reliability of the prediction.

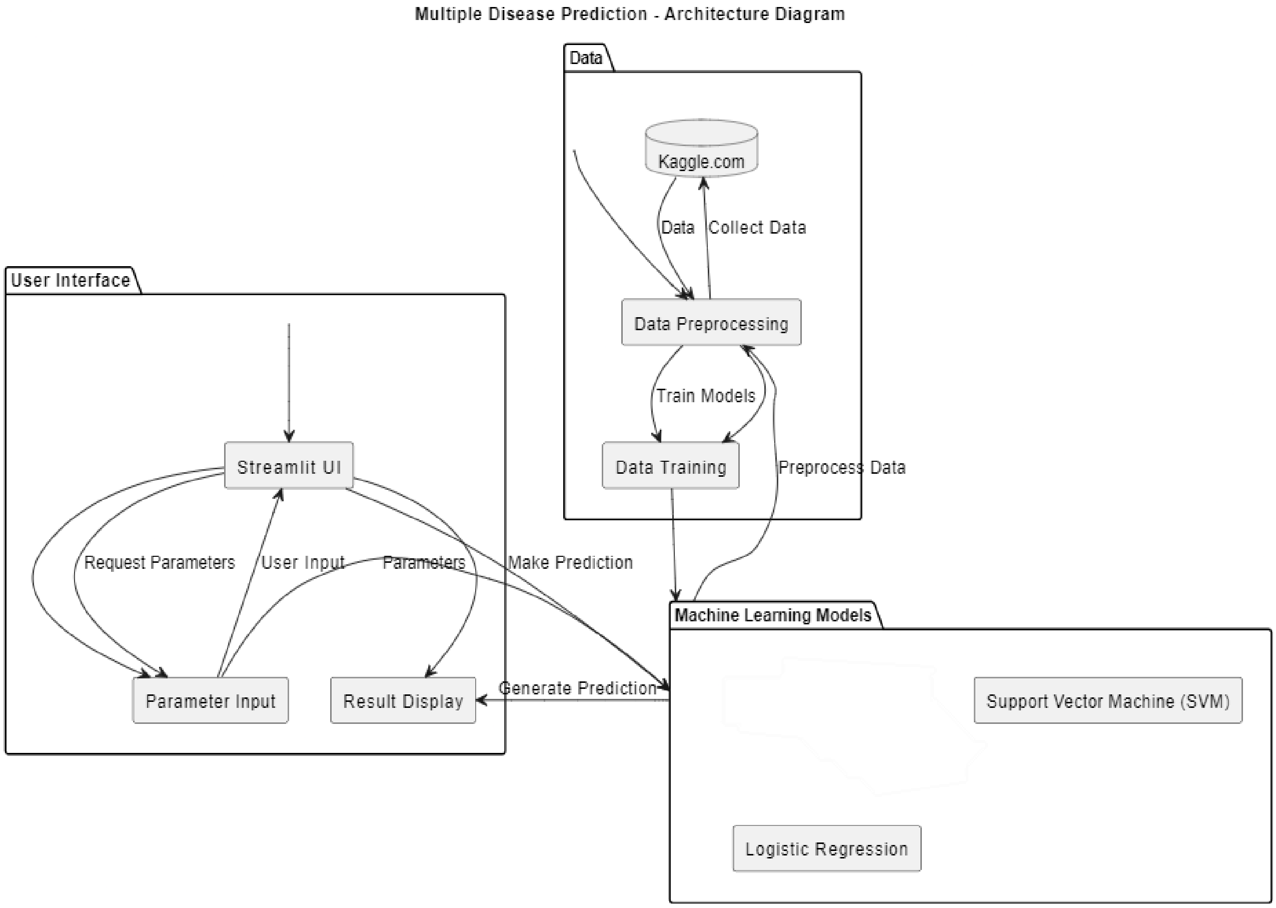
Output Design:

Once the necessary inputs are submitted, the system processes the data through the trained machine learning model and provides a clear prediction result. This result is displayed directly on the interface in an easy-to-understand message. For instance, the application may show:
"Prediction: The person is likely affected by [Disease Name]."
"Prediction: The person is not likely affected by [Disease Name]."
These messages are designed to be straightforward so users can easily understand the outcome. The clean presentation of results supports informed decision-making and encourages users to consult a medical professional when necessary. Together, the input and output design aim to create a seamless and user-friendly experience, helping individuals assess potential health risks with minimal effort.

VII. SYSTEM DESIGN

SYSTEM ARCHIETECTURE:

The Architecture diagram for the multiple disease prediction web application:



VIII. RESULT

The results for all the ML models and of final completed project are shown in the following figures and tables:

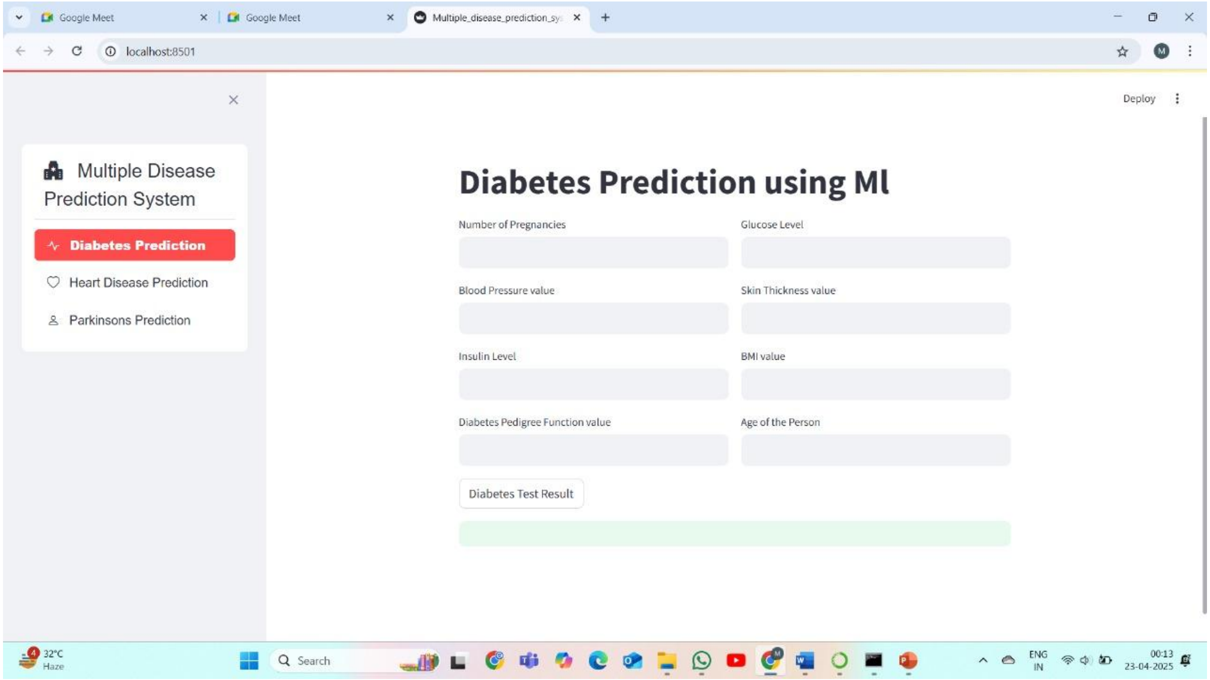
Table 1. Comparison of Accuracy of all 3 models

SN.	Disease Name	Algorithm Name	Existing system accuracy	Proposed system accuracy
1	Diabetes	SVM Classifier	76%	78%
2	Heart disease	Logistic Regression	80%	85%
3	Parkinson’s disease	SVM Classifier	71%	87%

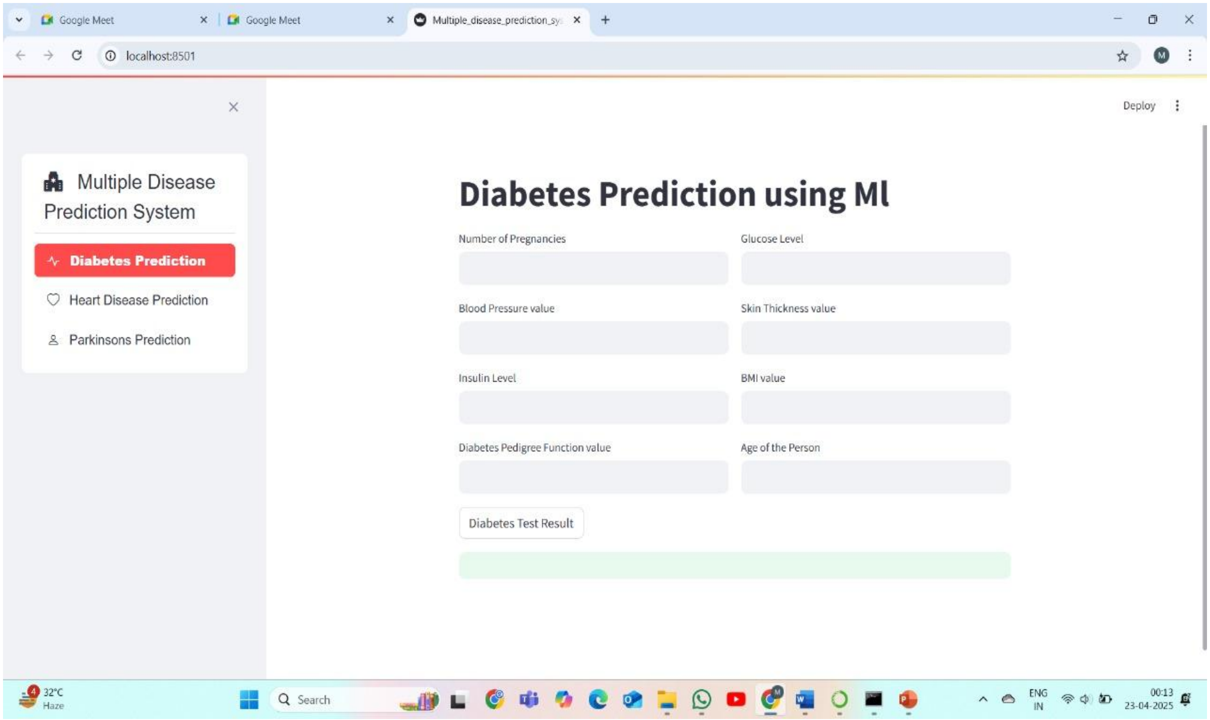
After completion of project the application interfaces are look like following pictures:

For diabetes and heart disease the features are less so we easily enter feature values in the respective feature input.
For Parkinson's disease the features are more so the application takes features values in a single input field, the values must be separated by comma (”,”).

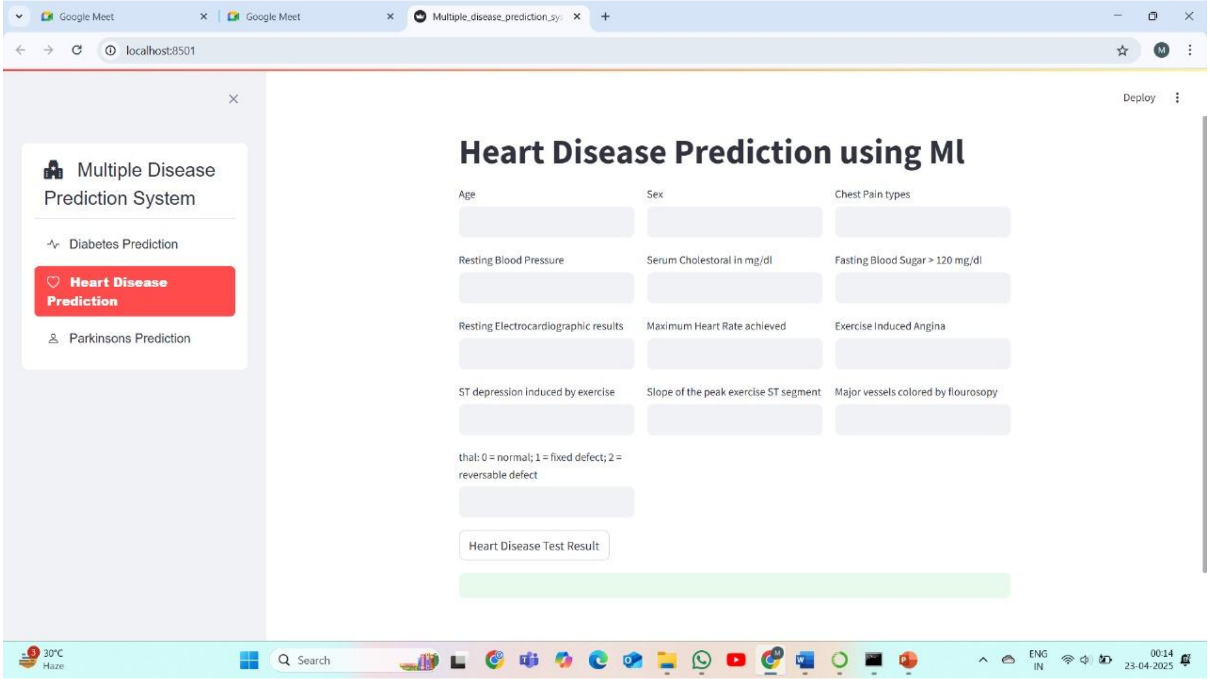
1. HOME PAGE



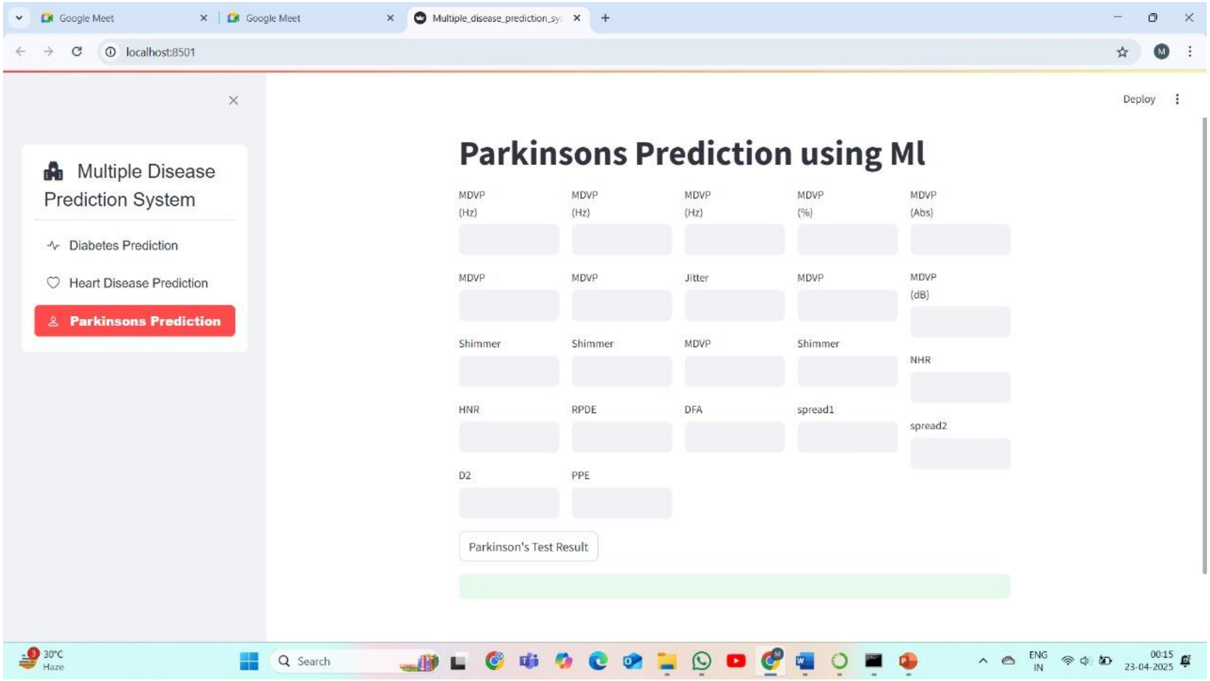
2.DIABETES PREDICTION



3.HEART DISEASE PREDICTION:



4.PARKINSON'S DISEASE PREDICTION:



IX. CONCLUSION

To sum up, this project demonstrates how machine learning can be effectively applied to predict multiple diseases through a simple and interactive web application. We focused on three major health conditions—diabetes, heart disease, and Parkinson's disease—using models built with Support Vector Machine (SVM) and Logistic Regression. The datasets used were collected from Kaggle, and thorough preprocessing was carried out to ensure reliable input for model training.

The performance of each model was promising. The SVM classifier achieved 78% accuracy in predicting diabetes and 89% accuracy for Parkinson's disease. Meanwhile, Logistic Regression gave an 85% accuracy rate for heart disease prediction. These results reflect the models' ability to make accurate assessments when provided with proper data.

The application is built with Streamlit, offering a user-friendly experience where individuals can choose a disease, input their health parameters, and instantly receive prediction results. While these models have shown strong performance, it's essential to recognize that results might vary with different datasets and model refinements.

In essence, this project highlights the growing role of machine learning in healthcare, especially in early diagnosis and risk assessment. With further improvements—such as the addition of more diseases, advanced algorithms, or real-time data monitoring—this system can evolve into an even more powerful tool for both healthcare providers and the general public.

X. FUTURE SCOPE

The "Multiple Disease Prediction using Machine Learning and Streamlit" project has demonstrated strong performance in predicting conditions like diabetes, heart disease, and Parkinson's disease. While the current results are promising, there are several exciting directions for future improvements and growth:

Broader Disease Coverage: At present, the system focuses on just three diseases. By expanding the range to include other health conditions, we can make the system more versatile and beneficial for a wider range of users.

Exploring Additional Machine Learning Algorithms: Currently, we are using algorithms like Support Vector Machines (SVM) and Logistic Regression. To boost model accuracy, it might be worthwhile to experiment with other algorithms such as Random Forests, Gradient Boosting, or Neural Networks. This could enhance the model's overall performance and reliability.

Advancing Feature Engineering: Feature engineering is critical for improving the quality of our predictions. By adopting more advanced methods such as dimensionality reduction, feature selection, and feature extraction, we can refine our models further, improving both accuracy and the ability to explain results.

Real-time Health Monitoring and Feedback: One key area for future development is introducing real-time monitoring. Features such as personalized health reminders, suggestions for disease prevention, and alerts when health parameters fall outside normal ranges could significantly enhance user experience and empower individuals to take better control of their health.

Implementing Explainable AI: To build trust and ensure transparency, it's essential to make the machine learning models more interpretable. Focusing on techniques for explainable AI, like feature importance visualization, can help users better understand how predictions are made and feel more confident in the system's advice.

REFERENCES

[1] Support Vector Machine (SVM): Corinna Cortes and Vladimir Vapnik (1995). Support-vector networks. Machine Learning, 20(3), 273-297.

[2] Logistic Regression: Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). John Wiley & Sons.

[3] Streamlit: Streamlit Documentation. https://docs.streamlit.io/

[4] Kaggle: Kaggle website. https://www.kaggle.com/

[5] Data sources: You can provide the specific datasets you used from Kaggle.com, mentioning the authors or contributors of the datasets.

[6] Zhang, Y., & Ghorbani, A. (2019). A review on machine learning algorithms for diagnosis of heart disease. IEEE Access, 7, 112751-112760.

[7] Arora, P., Chaudhary, S., & Rana, M. (2020). Prediction of diabetes using machine learning algorithms: A review. Journal of Ambient Intelligence and Humanized Computing, 11(6), 2575-2589.

[8] Saeed, A., & Al-Jumaily, A. (2020). Machine learning techniques for Parkinson's disease diagnosis using handwriting: A review. Computers in Biology and Medicine, 122, 103804.