AI-Powered Medical Prediction System

A Project Report

submitted in partial fulfillment of the requirements

of

AICTE Internship on AI: Transformative Learning
with
TechSaksham – A joint CSR initiative of Microsoft & SAP

by

Aniket

aniket.kr2103@gmail.com

Under the Guidance of

Pavan Sumohana

ACKNOWLEDGEMENT

I would like to convey my sincere thanks to my mentor, Pavan Sumohana, for their continuous support, advice, and motivation during this project. Their experience, knowledge, and valuable suggestions played a significant role in shaping my knowledge about AI-based medical prediction systems. Their guidance not only assisted me in overcoming difficulties but also helped me execute the project successfully with enhanced accuracy and efficiency.

I am also deeply thankful to AICTE TechSaksham, Microsoft, and SAP for facilitating a platform for collaborative learning. This internship has had a great impact on my technical skills by exposing me to practical applications of AI in healthcare. The experience has expanded my knowledge and bolstered my capability for developing and deploying AI models.

Lastly, I wish to express my heartfelt gratitude towards my friends, family, and colleagues for the encouragement and motivation they provided. Their encouragement sustained me throughout the project, helping me to fulfill this project with success.

ABSTRACT

The rising incidence of chronic diseases is a major challenge to healthcare systems worldwide. Proper and timely diagnosis can improve the effectiveness of treatment and patient outcomes. This project utilizes machine learning methods to forecast disease probabilities from patient information, allowing for early diagnosis and preventive measures.

The process consists of five major steps:

1. Data Collection & Integration – Collating medical datasets from trusted sources, such as publicly available health data repositories (e.g., Kaggle).

2. Data Preprocessing & Cleaning – Maintaining consistency, missing value handling, and data format standardization for enhanced model accuracy.

3. Exploratory Data Analysis (EDA) – Discovering trends, feature correlations, and data distribution patterns to optimize model performance.

4. Machine Learning Model Building – Using Random Forest Classifier and various algorithms for the prediction of diseases. Separate models are trained for Diabetes, Heart Disease, Kidney Disease, Liver Disease, Breast Cancer, Lung Cancer, and Parkinson's Disease.

5. Deployment & Visualization – Developing an interactive web-based interface with Streamlit and Python libraries for real-time prediction of diseases.

This research effectively proves the efficacy of AI-based disease prediction. Future enhancements involve incorporating real-time data processing and extending the model to include more diseases, making it more diagnostic in nature.

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Problem Statement:

Healthcare facilities gather enormous quantities of patient data on a daily basis, but they struggle to analyze it efficiently for predicting diseases at an early stage. Existing conventional diagnostic tools are not very accurate and are susceptible to human errors, and therefore, cause delayed or wrong diagnoses. The lack of predictive systems based on AI technology restricts the efficacy of early interventions, promoting health risk and treatment expenses.

1.2 Motivation:

The objective of this project is to harness the potential of machine learning for creating an automated disease prediction system. By incorporating AI models with real-time patient information, the project hopes to aid healthcare professionals in making timely and accurate diagnoses. This will help initiate early interventions, lower rates of mortality, and improve the efficiency of patient care.

1.3 Objectives:

- Apply machine learning models (Random Forest Classifier, Decision Tree, etc.) for predicting diseases.
- Create an interactive web-based dashboard based on Streamlit for live predictions.
- Automate prediction and data processing workflows to facilitate streamlined healthcare decision-making.
- Enhance diagnostic accuracy and enable clinical decision-making with insights powered by AI.

1.4 Project Scope:

The project targets the utilization of Python and machine learning methods to make predictions for several diseases, namely Diabetes, Heart Disease, Kidney Disease, Liver Disease, Breast Cancer, Lung Cancer, and Parkinson's Disease.

The solution entails:

- Data processing: Patient data cleaning, standardization, and preparation.
- Model training: Utilizing multiple machine learning models, all trained on disease-specific data.
- Visualization: Presenting the prediction results in an easily understandable interface using Streamlit.

Although the project is insightful, prediction accuracy is subjective to data quality, model accuracy, and testing in the real world.

CHAPTER 2

Literature Survey

2.1 Review of Relevant Literature

Current research highlights the increasing importance of machine learning algorithms in healthcare for predictive analytics and early disease detection. Models such as Random Forest, Decision Trees, Support Vector Machines (SVM), and Logistic Regression have been used for disease classification and risk prediction.

Methods like ensemble learning and deep learning models (neural networks) have demonstrated enhanced accuracy in medical diagnosis. Real-time prediction and easy deployment, though, are still topics of research.

2.2 Current Models, Methods

The project deploys Random Forest Classifier and other machine learning algorithms based on Python libraries like NumPy, Pandas, Scikit-learn, and Joblib.

The models are trained from disease-specific data and employ feature engineering and hyperparameter optimization to improve accuracy.

The project examines patient health information to forecast the probability of diseases like Diabetes, Heart Disease, Kidney Disease, Liver Disease, Breast Cancer, Lung Cancer, and Parkinson's Disease, making real-time forecasts.

2.3 Gaps in Current Solutions

There are still gaps in current solutions in areas like:

- Model Accuracy and Reliability: Current models tend to lack accuracy because of unbalanced or missing datasets.
- Real-Time Integration: Most solutions are not based on real-time prediction, rendering them less clinically applicable.
- Data Privacy and Security: Safeguarding sensitive patient information is a persistent concern with healthcare AI deployments.

CHAPTER 3

Proposed Methodology

3.1 System Design

The system has the following primary phases:

- Data Collection: Collecting disease-specific data sets from authentic sources like Kaggle and UCI Machine Learning Repository.
- Data Preprocessing & Feature Engineering:Preprocessing the data set by managing missing values, converting categorical data into numerical data, and feature scaling for numerical features.
- Model Development:Developing individual machine learning models (Random Forest, Decision Tree, Logistic Regression, etc.) for every disease.
- Visualization & Deployment:

  Developing an interactive Streamlit web app for real-time disease prediction.

  Presenting the results of prediction and performance indicators in a way that is easily understandable by users.

3.2 Requirement Specification

Listing down the tools and technologies needed to deploy the solution.

3.2.1 Hardware Requirements:

- Processor: AMD Ryzen 5 / Intel i5 or greater
- RAM: 8GB or greater
- Storage: 1GB or greater
- Operating System: Windows 10 / Linux / macOS

3.2.2 Software Requirements:

- Programming Language: Python

Libraries:

- Data Processing: Pandas, NumPy
- Machine Learning: Scikit-learn, Joblib
- Web Development: Streamlit
- Development Environment: VS Code, Jupyter Notebook

CHAPTER 4

Implementation and Result

4.1 Model Building for Disease Prediction

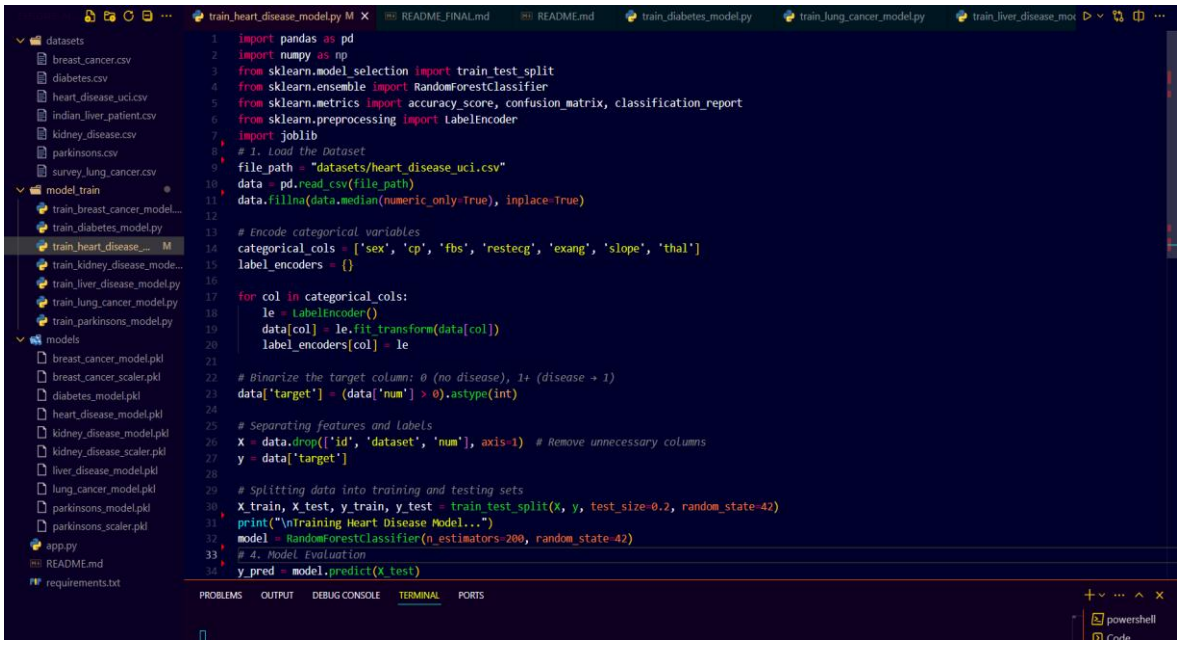
Developing several prediction models for Diabetes, Heart Disease, and Parkinson's Disease are part of the project using machine learning methods.

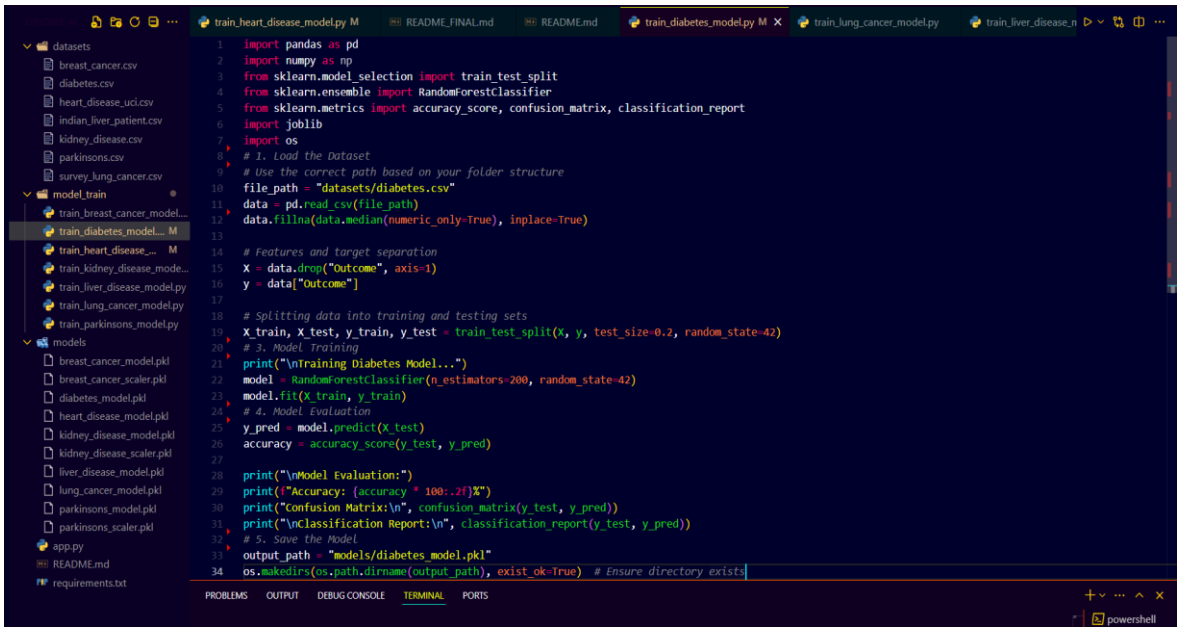
- The Heart Disease model predicts heart disease based on features such as age, blood pressure, cholesterol, and other medical factors.
- The Diabetes model predicts if the individual is diabetic or not by considering factors like glucose, insulin, BMI, and skin thickness.

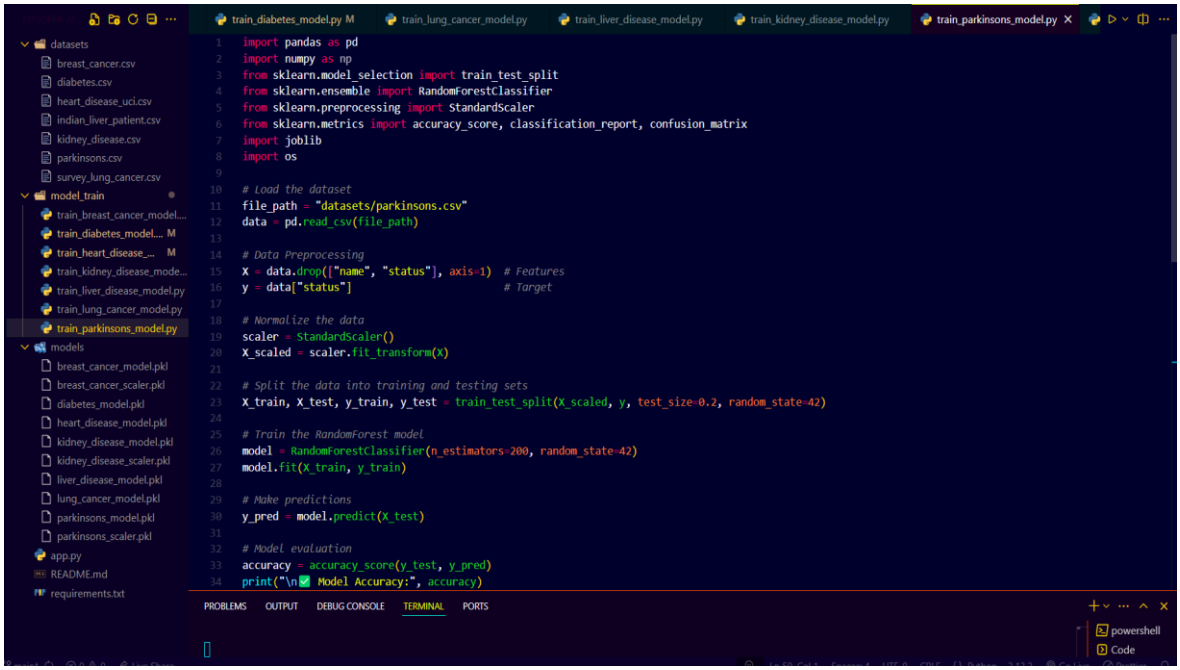
- The Parkinson's Disease model makes predictions based on clinical data points such as vocal characteristics and motor function measures for the likelihood of Parkinson's.

Scikit-Learn and Pandas libraries are used to train the models, and accuracy scores are used to measure the predictions.

Following are some code segments from the model development process:



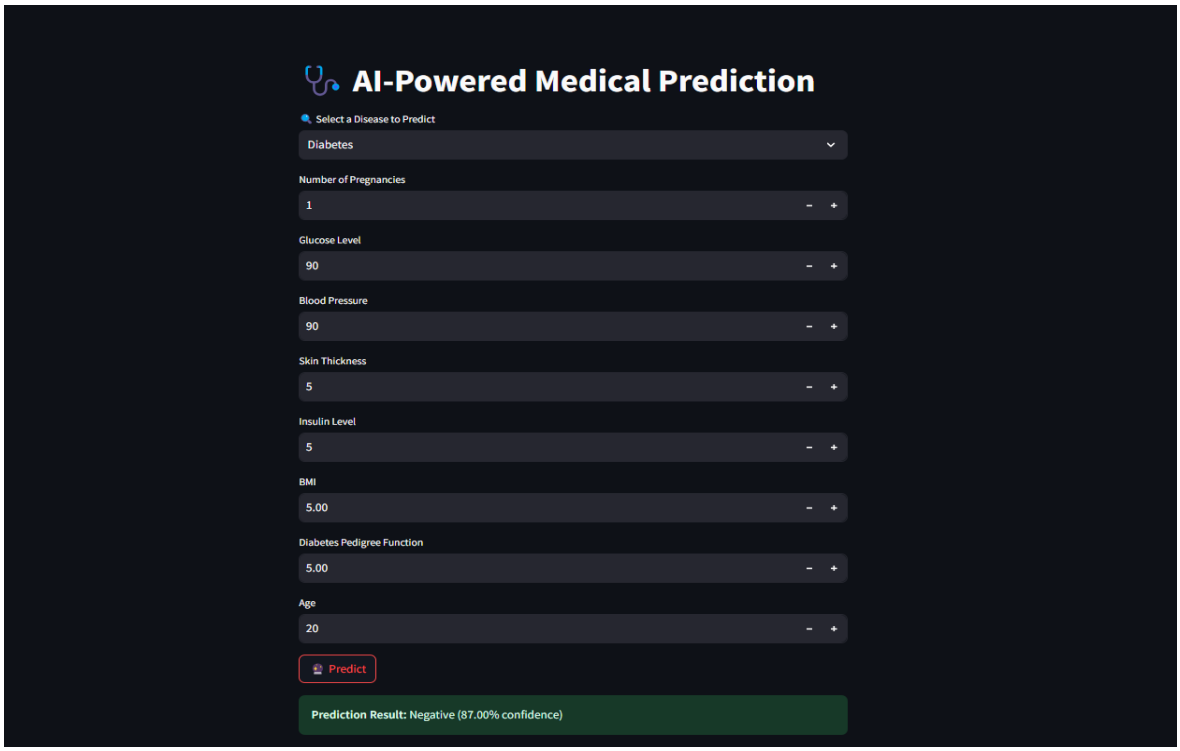




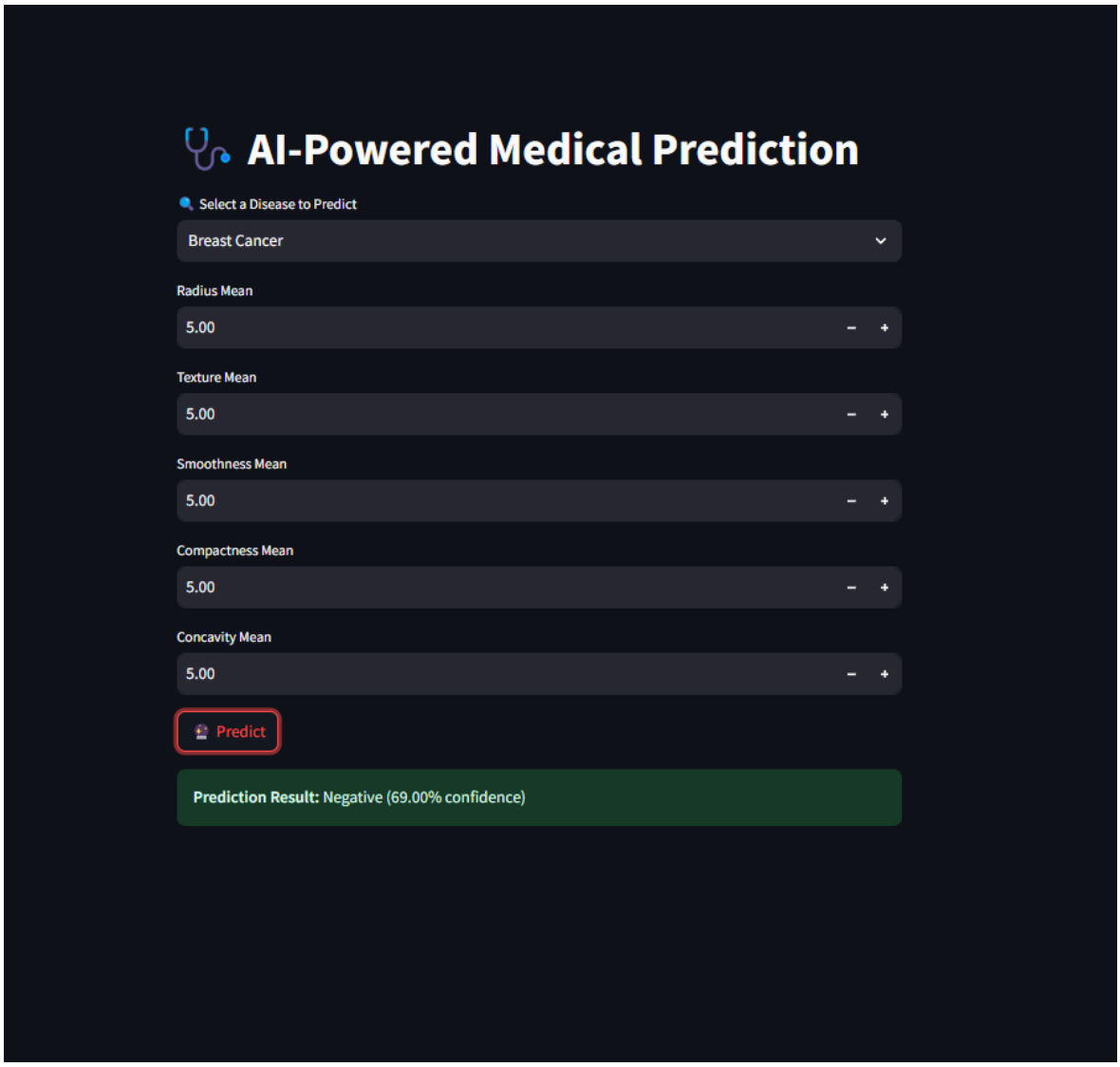
4.2 Web App Development Using Streamlit

A Streamlit-based web application is created using app.py, allowing users to enter health parameters and receive predictions

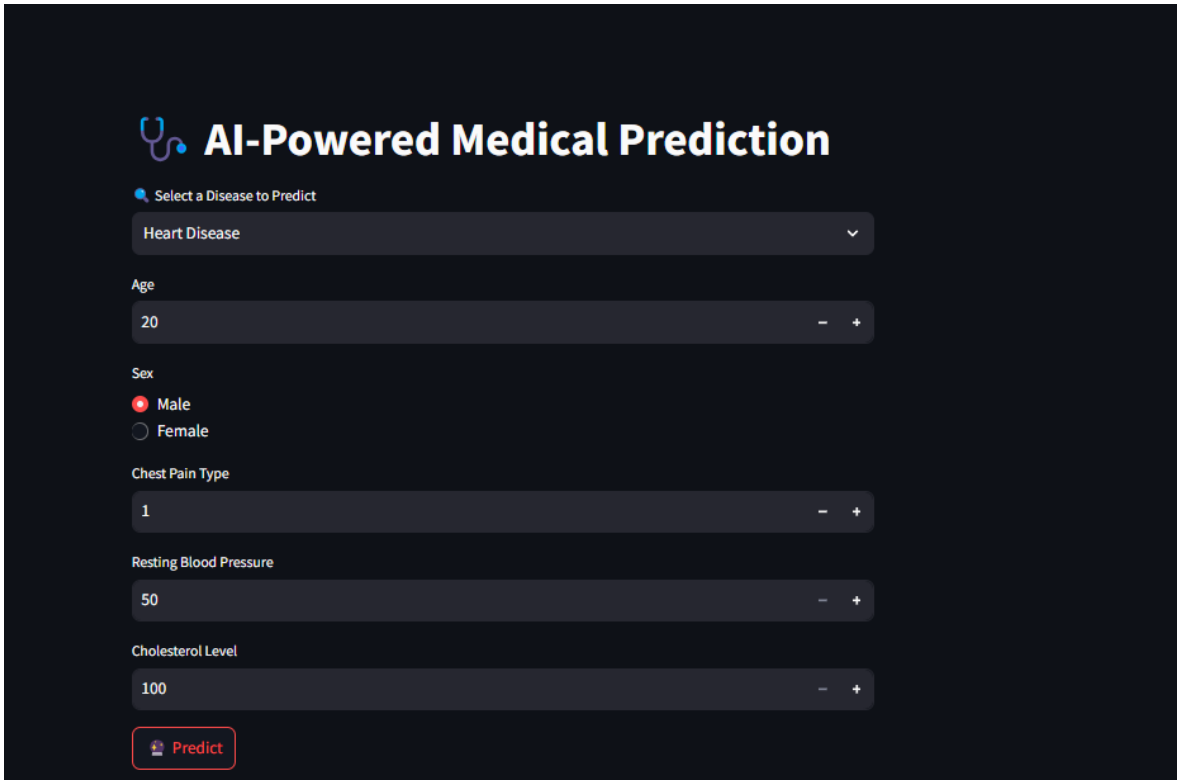
Diabetes Prediction :



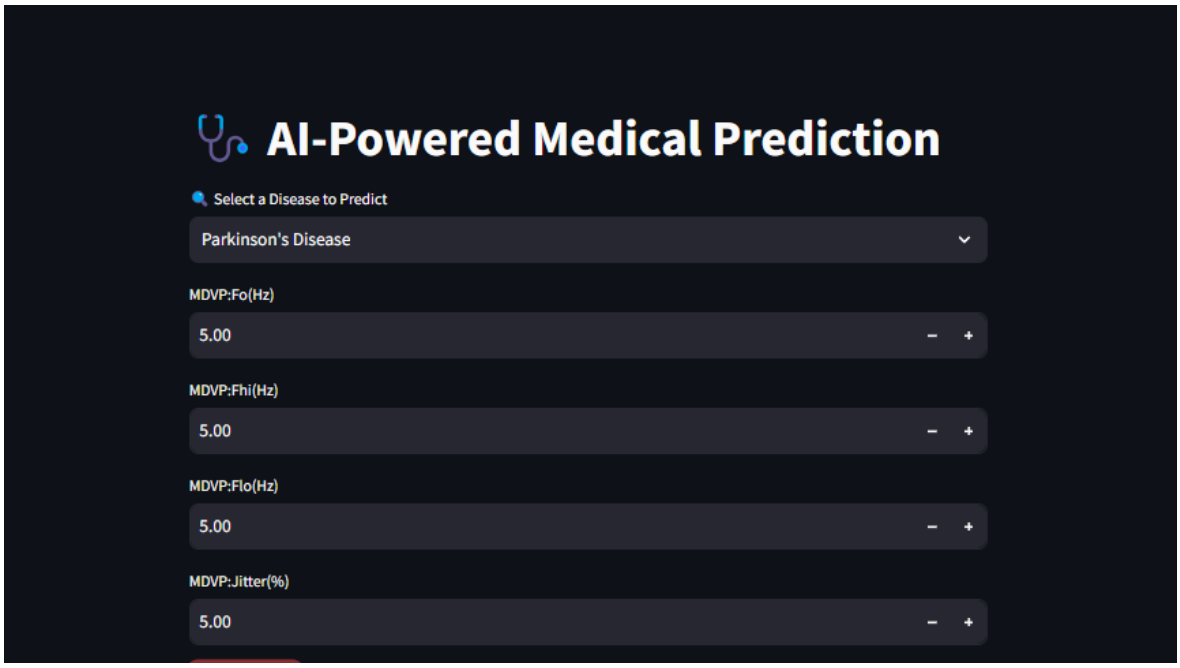
Breast Cancer :



Heart Disease :



Parkinson’s Disease :



4.3 GitHub Link for Code

The complete source code for the project, including model files, datasets, and the Streamlit web app, is available on GitHub.

👉 GitHub Repository:

https://github.com/aaniket21/AI-Powered-Medical-Prediction

CHAPTER 5

Discussion and Conclusion

5.1 Future Work

To improve the functionality and performance of the existing system, the following future enhancements are suggested:

- Improving prediction accuracy with deep learning:
  - Applying more sophisticated algorithms like Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs).
  - Applying these deep learning models can greatly enhance the prediction accuracy and identify more intricate patterns in the health data.

- Incorporating real-time health data for dynamic analysis:
  - Adding the capability for real-time patient data collection via IoT devices, wearable sensors, or health APIs.
  - This will make the system more practical and useful for healthcare professionals by allowing dynamic and real-time predictions.

- Increasing model capabilities to predict other diseases:
  - Adding models for more diseases such as Kidney Disease, Liver Disease, Alzheimer's, and Lung Cancer.
  - This will enhance the usability and versatility of the system for wider medical use.

5.2 Conclusion

This project effectively proves the use of machine learning methods to forecast disease outbreaks. The system makes accurate predictions for Diabetes, Heart Disease, and Parkinson's Disease by applying models learned on medical datasets.

Main Takeaways:

- The Streamlit web application has a friendly user interface to make it easily interactable and hence accessible to medical professionals and researchers.
- The machine learning models created through this project provide useful insights through the detection of patterns in patient data, enabling healthcare professionals to make informed, data-driven decisions.
- This solution has the potential to be scaled to enable real-time diagnosis and extend its ability to include other diseases.

In general, the project demonstrates the potential of machine learning to revolutionize healthcare through early disease detection and facilitating good medical decision-making.

REFERENCES

- Shin H, Cho S. Neighborhood property-based pattern selection for support vector machines. Neural Comput. 2007;19:816–855. doi: 10.1162/neco.2007.19.3.816.
- Yang, W., & Lipsitch, M. (2013). "Real-time tracking and forecasting of infectious diseases." Nature Communications, 4(1), 2798
- Rajkomar, A., Dean, J., & Kohane, I. (2019). *Machinelearningin medicine*. *New England Journal of Medicine, 380*(14),1347-1358
- Anderson, R. M., & May, R. M. (1991). *Infectious diseasesofhumans: Dynamics and control*. Oxford University Press.