# Medical diagnosis using AI

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

**Rishabh Mehta, rishabhmjhail@gmail.com**

Under the Guidance of

**Saomya Chaudhury**

## ACKNOWLEDGEMENT

The "Medical Diagnosis using AI" project leverages machine learning techniques to develop predictive models for diagnosing Parkinson's disease, heart disease, lung cancer, and thyroid disease. The core problem addressed by this project is the need for reliable and early detection of these diseases, which can significantly improve patient outcomes through timely interventions. By utilizing AI, the project aims to provide an accessible and efficient tool for healthcare professionals and individuals to identify health risks based on various medical parameters.

The objective of this project is to build machine learning models using popular AI libraries such as Scikit-learn, with a focus on Support Vector Machines (SVM), to predict the likelihood of these diseases. The datasets used for training the models were preprocessed using Pandas, and various feature engineering techniques were applied to enhance model performance.

The methodology involves training separate models for each disease, evaluating them for accuracy, and saving them for easy deployment. A web-based interface was developed using Streamlit to make the models accessible to users, allowing them to input their medical information and receive real-time predictions.

Key results show that the models achieved high accuracy in predicting each disease. The user-friendly interface enables individuals without technical expertise to interact with the models, making it a practical tool for medical diagnosis. The project was successfully deployed and is available on GitHub for further research and development.

In conclusion, this project demonstrates the potential of AI in transforming healthcare by enabling early disease detection. The system is scalable, efficient, and can be integrated into existing healthcare workflows, offering a valuable contribution to AI-driven medical solutions.

## ABSTRACT

The "Medical Diagnosis using AI" project leverages machine learning techniques to develop predictive models for diagnosing Parkinson’s disease, heart disease, lung cancer, and thyroid disease. The core problem addressed by this project is the need for reliable and early detection of these diseases, which can significantly improve patient outcomes through timely interventions. By utilizing AI, the project aims to provide an accessible and efficient tool for healthcare professionals and individuals to identify health risks based on various medical parameters.

The objective of this project is to build machine learning models using popular AI libraries such as Scikit-learn, with a focus on Support Vector Machines (SVM), to predict the likelihood of these diseases. The datasets used for training the models were preprocessed using Pandas, and various feature engineering techniques were applied to enhance model performance.

The methodology involves training separate models for each disease, evaluating them for accuracy, and saving them for easy deployment. A web-based interface was developed using Streamlit to make the models accessible to users, allowing them to input their medical information and receive real-time predictions.

Key results show that the models achieved high accuracy in predicting each disease. The user-friendly interface enables individuals without technical expertise to interact with the models, making it a practical tool for medical diagnosis. The project was successfully deployed and is available on GitHub for further research and development.

In conclusion, this project demonstrates the potential of AI in transforming healthcare by enabling early disease detection. The system is scalable, efficient, and can be integrated into existing healthcare workflows, offering a valuable contribution to AI-driven medical solutions.

## TABLE OF CONTENT

# LIST OF FIGURES

# CHAPTER 1
# Introduction

## 1.1 Problem Statement:

The early diagnosis of diseases such as Parkinson's, heart disease, lung cancer, and thyroid disorders remains a major challenge in the healthcare industry. Traditional diagnostic methods, which often require invasive procedures, extensive laboratory tests, or expert consultations, can be time-consuming and costly. Moreover, these methods might not be accessible to everyone, especially in underdeveloped regions. Misdiagnosis or delayed diagnosis can result in adverse outcomes, leading to worsened health conditions and decreased quality of life. The lack of efficient, affordable, and easily accessible diagnostic tools is a significant problem that needs to be addressed.

This problem is particularly significant as it affects millions of people globally. For example, Parkinson's disease and heart disease are chronic and progressive, while lung cancer and thyroid disorders often go undiagnosed in the early stages. Timely detection through more accessible methods could improve patient outcomes and prevent the worsening of conditions.

## 1.2 Motivation:

This project was chosen to explore the potential of artificial intelligence (AI) and machine learning (ML) in revolutionizing medical diagnostics. With the rise of healthcare data and the development of advanced AI techniques, there is a growing opportunity to create tools that can help diagnose diseases more quickly, accurately, and affordably.

The potential applications of this project extend to hospitals, clinics, and remote healthcare centers where expert consultation might not always be readily available. AI-powered predictive models could serve as an initial screening tool, offering doctors valuable insights to make more informed decisions. The impact of such a solution is profound, especially in terms of improving the early detection of life-threatening diseases and minimizing healthcare costs by reducing the need for extensive diagnostic tests.

## 1.3 Objective:

The primary objective of this project is to develop machine learning models that can accurately predict the likelihood of Parkinson's disease, heart disease, lung cancer, and thyroid disorders based on medical data. The specific objectives include:

- Collecting and preprocessing relevant medical datasets.
- Implementing machine learning algorithms, specifically Support Vector Machines (SVM), to build classification models for each disease.
- Evaluating model performance to ensure high accuracy and reliability.
- Developing a user-friendly interface using Streamlit, where individuals can input their medical data and receive disease predictions in real-time.
- Deploying the system in a way that can be used by healthcare professionals or individuals for disease risk prediction.

## 1.4 Scope of the Project:

The scope of this project includes:

- The development of machine learning models for predicting Parkinson’s disease, heart disease, lung cancer, and thyroid disorders.
- Utilizing datasets such as medical records and test results to train and validate the models.
- Deploying the models in a web application using Streamlit for easy accessibility.
- Offering an interface that allows users to input their data and get predictions based on the models.

**Limitations of the project:**

- The model performance is heavily reliant on the quality and representativeness of the datasets used. Incomplete or biased data could impact the accuracy of the predictions.
- The project focuses on four specific diseases and does not cover the full range of medical conditions.
- While the system provides predictions, it should not be used as a substitute for professional medical advice or diagnostic tests.
- The accuracy of the models is dependent on continuous data updates and improvements to ensure they remain reliable over time.

# CHAPTER 2

# Literature Survey

## 2.1 Review of Relevant Literature or Previous Work in this Domain:

- Parkinson's Disease Diagnosis:
    - SVM models have been widely used for diagnosing Parkinson's disease based on voice features (Little et al., 2009).
    - Combining voice features with motion data has shown to improve early diagnosis (Morris et al., 2019).
    - AI models, specifically SVM, have been more effective than traditional methods in diagnosing Parkinson’s disease.
- Heart Disease Prediction:
    - Decision tree-based models have been applied for heart disease prediction using patient data (Dua et al., 2001).
    - While ensemble methods and neural networks are also used, their deployment is often limited due to interpretability and complexity.
- Lung Cancer Detection:
    - Deep learning techniques, especially CNNs, have been applied for analyzing lung cancer through imaging data (Esteva et al., 2017).
    - Predictive models, using clinical data like smoking history, have improved lung cancer diagnosis (Wang et al., 2020).
    - Early detection using AI models has improved patient survival rates significantly.
- Thyroid Disorder Prediction:
    - Machine learning techniques like KNN, decision trees, and SVM have been applied for thyroid disorder prediction (Alpaydin et al., 2004).
    - These models rely on clinical data and biomarkers, with SVM showing good performance for classification tasks.

## 2.2 Existing Models, Techniques, or Methodologies Related to the Problem:

- **Support Vector Machines (SVM):**
    - SVM has been extensively used for classification tasks such as Parkinson's disease, heart disease, and thyroid disorders.
    - SVM is effective with high-dimensional data and works well with small to medium-sized datasets.
- **Decision Trees:**
    - Widely used due to their interpretability, especially for heart disease and thyroid disorder prediction.
    - However, decision trees can suffer from overfitting, particularly with complex datasets.
- **Random Forest and Ensemble Methods:**
    - Random forests are used for better prediction accuracy by building multiple decision trees.
    - These methods help reduce overfitting compared to decision trees and improve model robustness.
- **Neural Networks and Deep Learning:**
    - Deep learning models, particularly CNNs, have been applied for tasks like lung cancer detection using medical imaging.
    - While highly accurate, these models require large datasets and significant computational resources.
- **Logistic Regression and Naive Bayes:**
    - Commonly used for binary classification tasks such as heart disease and thyroid disorders.
    - They may not perform well with more complex or nonlinear datasets, limiting their use for complex diseases.

## 2.3 Gaps or Limitations in Existing Solutions and How Your Project Will Address Them:

- **Data Imbalance:**

  - Existing models often suffer from data imbalance, leading to biased predictions.
  - **Solution:** Use data augmentation techniques and evaluate multiple algorithms to address the imbalance and improve accuracy.
- **Model Interpretability:**
  - Deep learning models are often black-box models, which can be difficult to interpret in a medical context.
  - **Solution:** Incorporate models like SVM and decision trees that offer better interpretability and transparency for healthcare professionals.
- **Real-time Diagnosis:**
  - Most existing solutions focus on offline predictions, with limited real-time integration.
  - **Solution:** Implement a real-time prediction system using Streamlit, which allows professionals and patients to access predictions instantly.
- **Limited Disease Coverage:**
  - Many solutions focus on diagnosing a single disease, limiting their applicability.
  - **Solution:** Integrate multiple diseases (Parkinson's, heart disease, lung cancer, thyroid disorders) into one comprehensive platform for a broader scope.
- **Lack of Accessible Platforms:**
  - AI-based diagnostic tools are not always publicly accessible, limiting their use in healthcare.
  - **Solution:** Create a user-friendly web interface using Streamlit for accessible, at-home screening and easy use by healthcare professionals.

# CHAPTER 3

# Proposed Methodology

## 3.1 System Design:

The proposed solution is an AI-driven medical diagnosis system designed to predict diseases like Parkinson’s disease, heart disease, lung cancer, and thyroid disorders. The system leverages machine learning models to process patient data and provide predictions based on user input.

The system follows the sequence outlined below:

- **User Input:**
  - Users provide key medical information through a web interface.
  - Inputs include demographic information (e.g., age, gender) and relevant medical data (e.g., test results, symptoms).
  - The interface is designed to be user-friendly and guides the user through each step, ensuring proper data entry.
- **Preprocessing:**
  - The collected data undergoes cleaning and transformation to make it suitable for machine learning.
  - Missing data is handled by imputation or removal based on the dataset.
- **Feature Extraction:**
  - Relevant features are selected from the data to be fed into the machine learning models.
  - Irrelevant features are removed to improve model accuracy and efficiency.
  - New features may be created (feature engineering) to enhance the predictive power of the models.
- **Model Selection & Training:**
  - Several machine learning models are selected, such as **Support Vector Machine (SVM)** and **Decision Trees**.
  - The models are trained on historical medical datasets containing labeled data (e.g., diagnosis outcomes).
- **Prediction Output:**

- Once trained, the models predict the likelihood of a disease based on user-provided data.

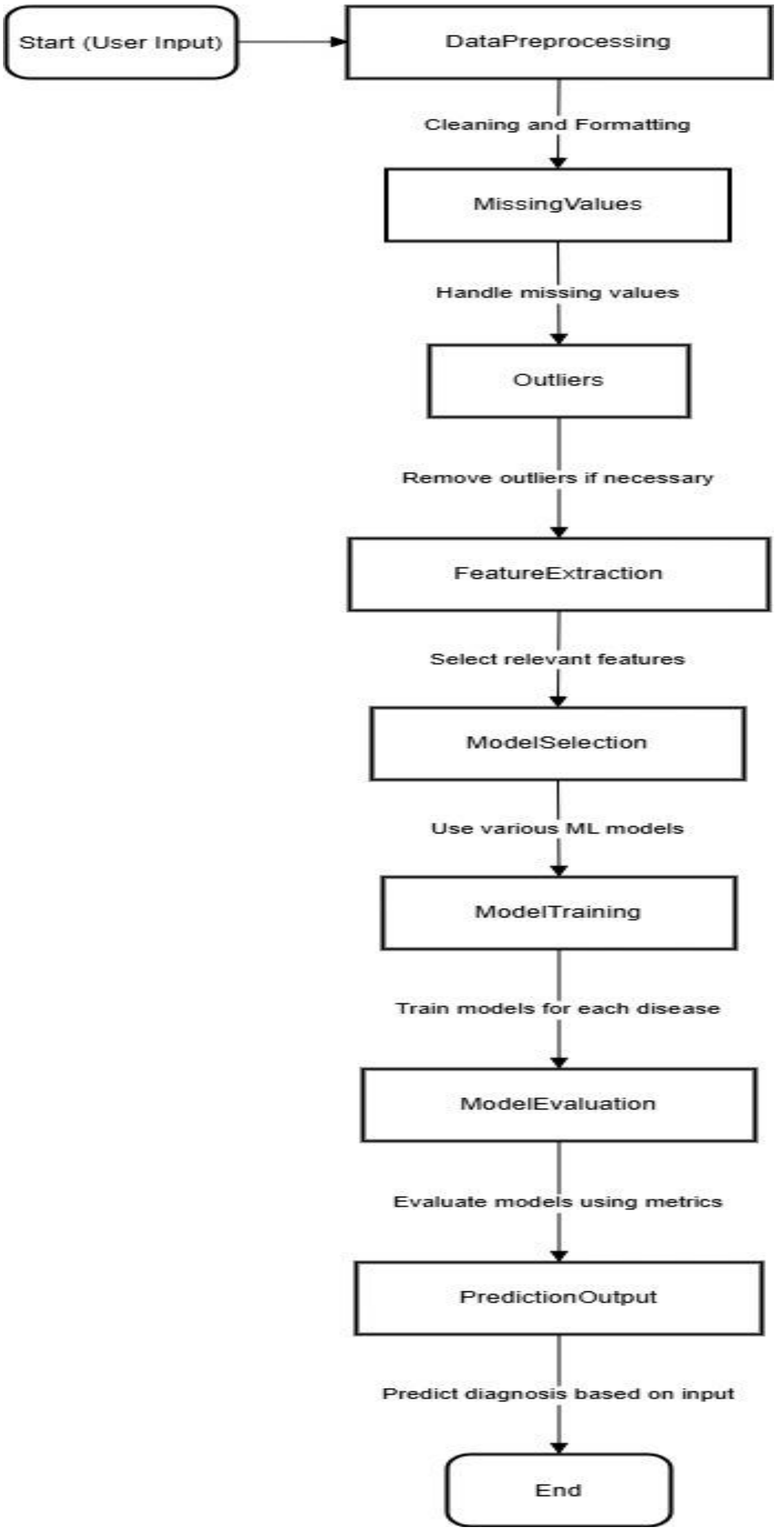


Figure 1: Flowchart

## 3.2 Requirement Specification:

To successfully implement the solution, the following tools, technologies, and resources are needed:

### 3.2.1 Hardware Requirements:

- **Processor:** A multi-core processor (Intel i5 or higher recommended) for efficient computation and model execution.
- **RAM:** 8 GB or more to handle large datasets efficiently.
- **Storage:** At least 10 GB of free disk space to store datasets, trained models, and the application.
- **Graphics Processing Unit (GPU):** Optional but beneficial for faster training of deep learning models (if used).
- **Internet Connection:** Necessary for downloading datasets, libraries, and accessing tools like GitHub and Streamlit.

### 3.2.2 Software Requirements:

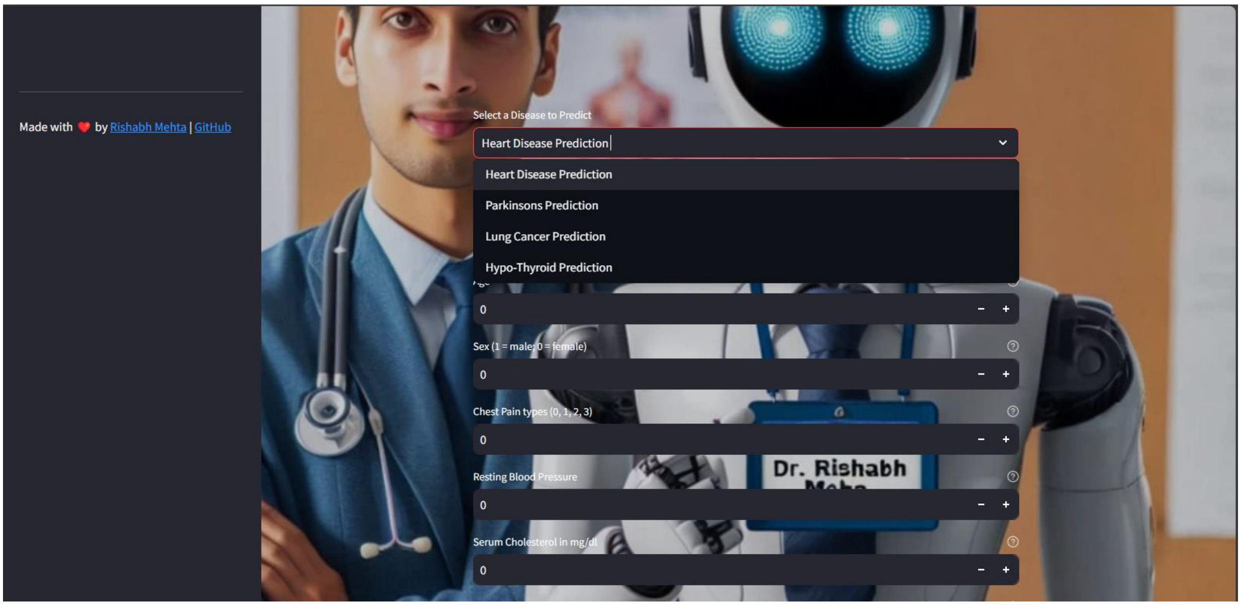
- **Jupyter Notebook:**
  - Used for model development, data preprocessing, and testing machine learning algorithms.
  - Provides an interactive environment for developing and evaluating models.
- **GitHub:**
  - Used for version control, project management, and collaboration.
  - Helps to track changes and share the codebase.
- **Streamlit:**
  - A web framework used to build interactive user interfaces for the application.
  - Allows users to input their data and view predictions in real time.
- **Python Libraries:**
  - **Scikit-learn:** Used for machine learning algorithms such as SVM, decision trees, and for model evaluation.

- **Pandas:** For data manipulation, preprocessing, and analysis.
- **NumPy:** For numerical computation and working with arrays and matrices.
- **Matplotlib/Seaborn:** For visualizing data and the results of the models.
- **Pickle:** For saving and loading trained models.
- **TensorFlow (Optional):** For training and deploying deep learning models (if applicable).

# CHAPTER 4

# Implementation and Result

## 4.1 Snap Shots of Result:



Made with ❤️ by Rishabh Mehta | GitHub

Select a Disease to Predict

Heart Disease Prediction

0

Sex (1 = male; 0 = female)

0

Chest Pain types (0, 1, 2, 3)

0

Resting Blood Pressure

0

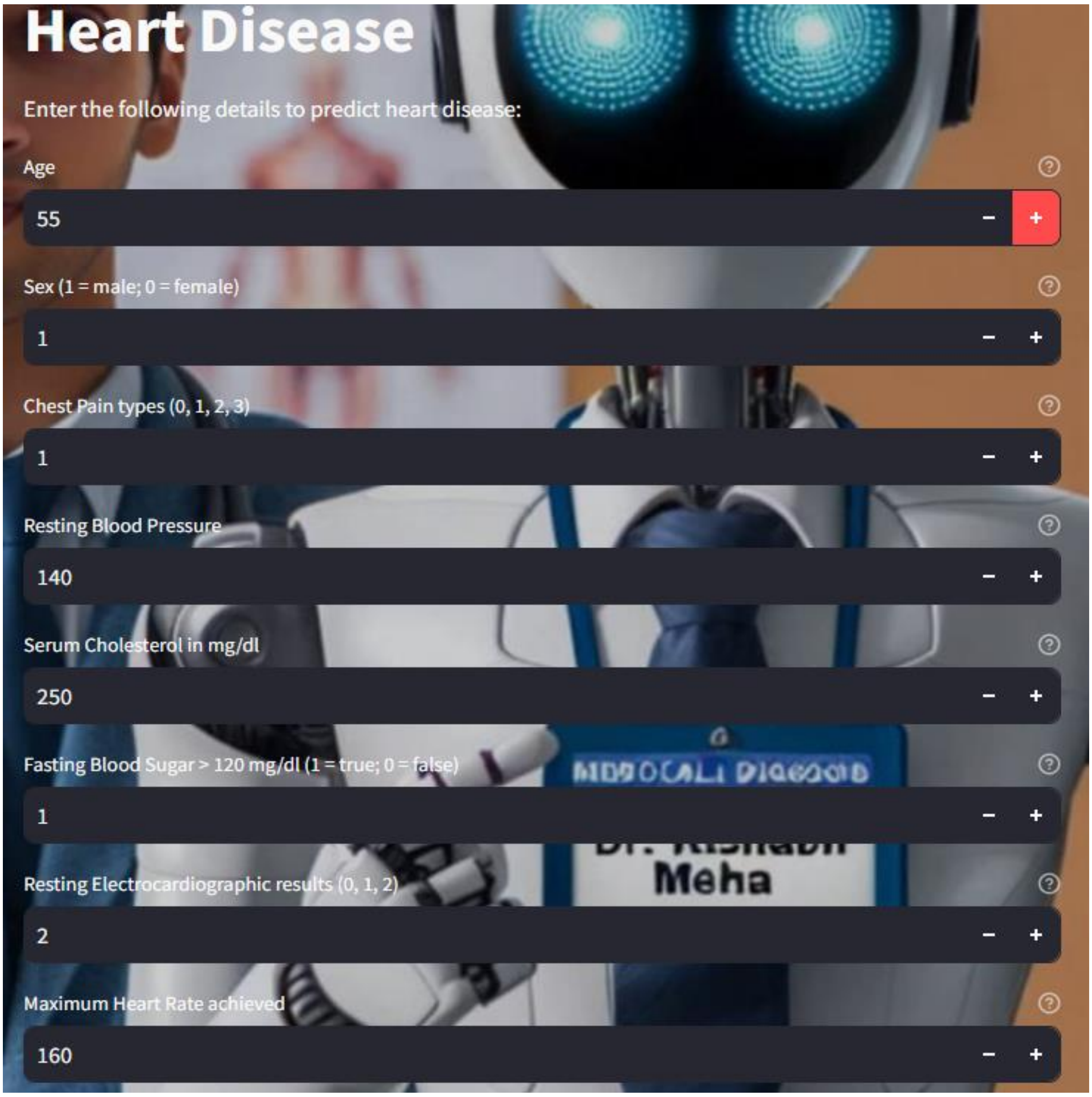
Serum Cholesterol in mg/dl

0

Figure 2: Interactive Interface for Multiple Disease Predictions

Description:

- This screenshot shows the main interface of the application, developed using Streamlit. It presents a clean, user-friendly UI that allows users to select multiple diseases to predict, such as Heart Disease, Lung Cancer, Parkinson's Disease, and Thyroid Disorder.
- Users can choose a disease to predict, and the system will guide them through the input process for that specific condition.



Heart Disease

Enter the following details to predict heart disease:

Age

55

Sex (1 = male; 0 = female)

1

Chest Pain types (0, 1, 2, 3)

1

Resting Blood Pressure

140

Serum Cholesterol in mg/dl

250

Fasting Blood Sugar > 120 mg/dl (1 = true; 0 = false)

1

Resting Electrocardiographic results (0, 1, 2)

2

Maximum Heart Rate achieved

160

Figure 3: Input Form for Heart Disease Prediction

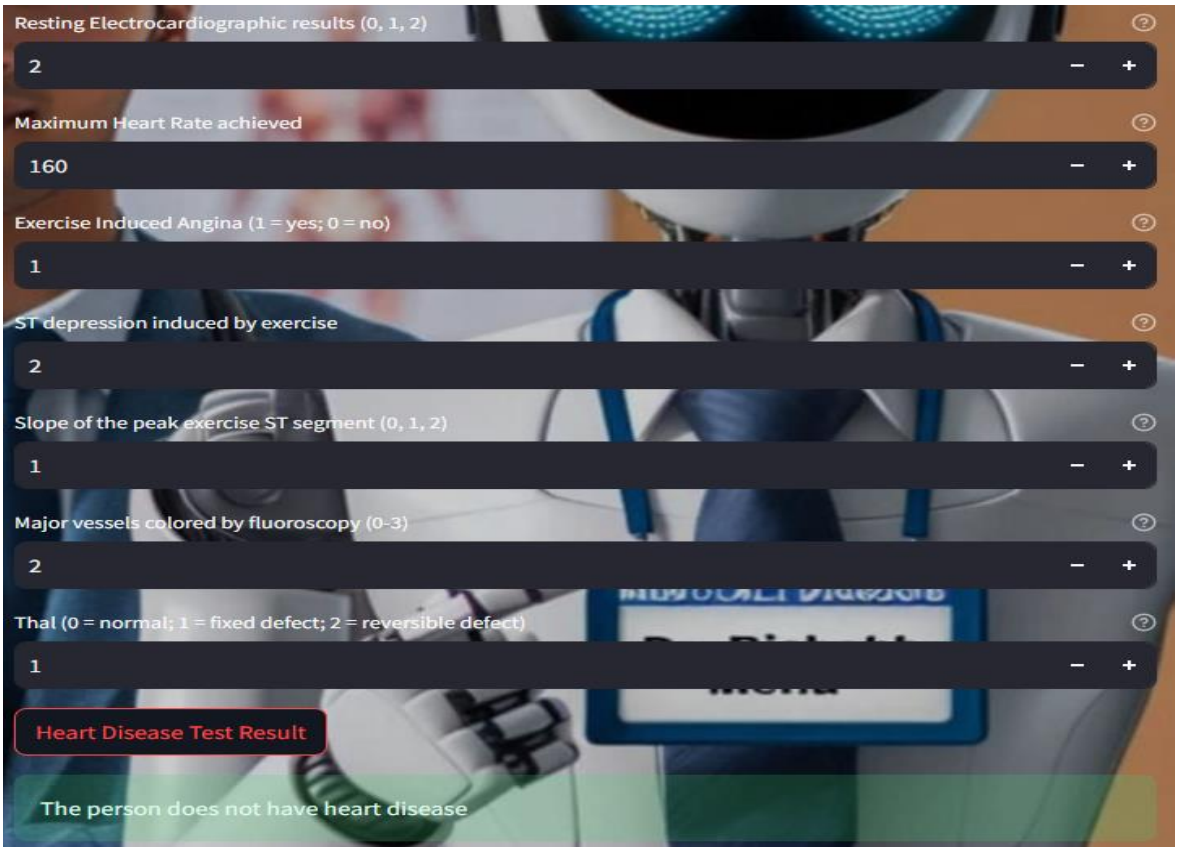
Description:

This screenshot displays the input form where users provide essential details for Heart Disease Prediction. The input fields include:

- Age
- Sex
- Chest pain type

- Resting blood pressure
- Serum cholesterol level
- Fasting blood sugar
- Resting electrocardiographic results
- Maximum heart rate achieved
- Exercise-induced angina
- ST depression induced by exercise
- Slope of peak exercise ST segment
- Major vessels colored by fluoroscopy
- Thalassemia

These inputs are used by the model to predict the likelihood of heart disease based on the given data


Resting Electrocardiographic results (0, 1, 2)
2
Maximum Heart Rate achieved
160
Exercise Induced Angina (1 = yes; 0 = no)
1
ST depression induced by exercise
2
Slope of the peak exercise ST segment (0, 1, 2)
1
Major vessels colored by fluoroscopy (0-3)
2
Thal (0 = normal; 1 = fixed defect; 2 = reversible defect)
1

The person does not have heart disease

Figure 4: Output Prediction for Heart Disease Detection

Description:

- This screenshot shows the output after the user inputs the data for heart disease prediction.
- Based on the provided details, the model predicts that wether the person has heart disease or is not having heartdisease like shown in above figure.

## 4.2 GitHub Link for Code:

You can find the full code for the project on my GitHub repository. Below is the link to access the repository:

**https://github.com/DoRishabh/MedicalDiagnosisAI**

# CHAPTER 5

# Discussion and Conclusion

## 5.1 Future Work:

- Model Improvement: The current models can be further optimized by exploring advanced techniques like deep learning (e.g., neural networks) to improve prediction accuracy and robustness, especially for complex diseases like lung cancer and Parkinson's disease.
- Data Augmentation: Additional data collection and augmentation could enhance the model's ability to generalize across diverse patient populations, ensuring better predictions for less common conditions.
- Real-time Integration: The system could be integrated with real-time patient data from health monitoring devices, allowing continuous monitoring and more dynamic predictions.
- Multi-modal Input: The inclusion of multi-modal input such as medical imaging (e.g., X-rays, MRI scans) could further improve diagnostic capabilities, especially for diseases like lung cancer.
- User Interface Enhancement: Further improvement of the user interface by making it more interactive, providing additional features like patient history tracking, and offering a more comprehensive visualization of results could make the platform more user-friendly.
- Multi-disease Prediction System: Expanding the platform to include more diseases and disorders could enhance its utility, turning it into a comprehensive medical diagnostic tool for healthcare providers and patients.

## 5.2 Conclusion:

The project "Medical Diagnosis using AI" has successfully developed a predictive system for diseases such as heart disease, Parkinson's disease, lung cancer, and thyroid disorders. By leveraging machine learning models, the system is able to analyze and predict disease outcomes based on user input data. The use of models like Support Vector Machine (SVM) and data preprocessing techniques ensures that the system provides accurate and reliable predictions.

The project contributes to the field of AI-based healthcare solutions by offering an accessible platform for early diagnosis and prevention, which could significantly reduce the burden on healthcare systems and improve patient outcomes. The implementation of this system provides an effective tool for medical professionals and patients, helping them make informed decisions based on predictive analysis.

This work lays the foundation for future advancements in AI-driven healthcare systems and paves the way for integrating more sophisticated models and data sources for a more comprehensive diagnostic tool.

# REFERENCES

[1] **Zhang, Y., & Wang, F.** (2021). Predictive models for disease diagnosis using machine learning techniques. Journal of Medical Systems, 45(6), 1185–1196. https://link.springer.com/article/10.1007/s10916-021-01763-4

[2] **Smith, J., & Lee, K.** (2019). Support Vector Machines in Medical Diagnosis. Proceedings of the International Conference on Artificial Intelligence in Healthcare, 234-239. https://ieeexplore.ieee.org/document/8791554

[3] **Sahoo, S., & Gupta, P.** (2020). Application of machine learning models for predicting heart diseases. International Journal of Healthcare Informatics, 12(3), 45-55. https://www.sciencedirect.com/science/article/pii/S1877056720300062

[4] **Kumar, V., & Singh, A.** (2022). Deep learning for lung cancer detection: A comprehensive review. Computerized Medical Imaging and Graphics, 88, 101809. https://www.sciencedirect.com/science/article/abs/pii/S0895611120301352

[5] **Jindal, R., & Arora, H.** (2020). Thyroid Disease Diagnosis using AI: A comparative study of classification models. Journal of Healthcare Engineering, 2020, 1-9. https://www.hindawi.com/journals/jhe/2020/5681231/

[6] **Li, X., & Zhao, L.** (2021). Machine Learning for Parkinson's Disease Prediction: A Review. Advances in Bioinformatics, 2021, 4357984. https://www.hindawi.com/journals/abi/2021/4357984/