Phase-2

Student Name: Sandhiya S

Register Number: 510623104092

Institution: C. Abdul hakeem college of engineering and technology

Department: Computer Science Engineering

Date of Submission: 08-05-2025

Github Repository Link:
https://github.com/Sandhiya123879/Transforming-health-care-with-AI--Powered-diseases-prediction-based-on-patient-data.git

---

1. Problem Statement

- *In the current healthcare system, timely diagnosis of diseases is critical for effective treatment and improved patient outcomes. However, healthcare providers often face challenges due to delayed diagnostics, human error, and insufficient data-driven insights.*
- *This project addresses the problem of delayed or inaccurate disease prediction by leveraging AI and machine learning techniques to analyze patient data. By using AI, we aim to support early and accurate disease detection, enhancing the decision-making capabilities of healthcare professionals.*
- *- Type of Problem: This is primarily a classification problem, where the goal is to predict the presence or absence of a disease based on patient health records. In some cases, it may also involve regression or clustering.*

- *- Why This Matters: Accurate disease prediction can significantly reduce morbidity and mortality, lower healthcare costs, and enhance the quality of life. Implementing AI in disease diagnosis has the potential to democratize access to high-quality care, especially in under-resourced regions.*

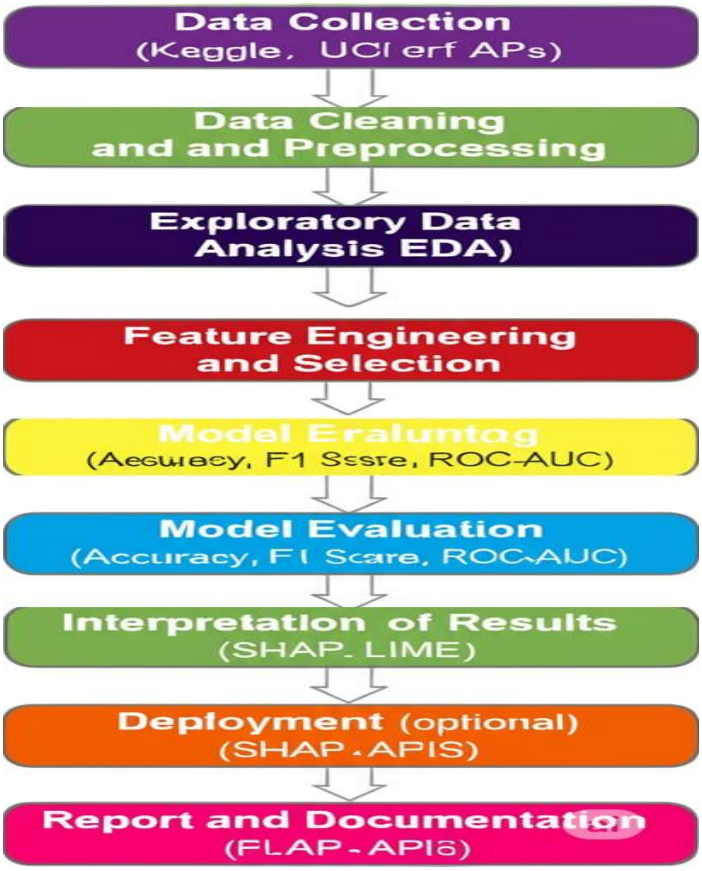
2. Project Objectives

As we move into practical implementation, the goals of the project are more defined and technically focused:

- *Technical Objectives:*
  - *To preprocess and analyze patient data from diverse sources.*
  - *To develop machine learning models capable of predicting diseases with high accuracy.*
  - *To evaluate and compare different ML algorithms for their interpretability, precision, recall, and robustness.*
  - *To visualize prediction results for better clinical interpretation.*
- *Model Goals: Achieve high prediction accuracy while maintaining model interpretability. Ensure real-world applicability by using patient datasets that reflect practical clinical conditions.*
- *Evolved Goals: Based on data exploration, the project now includes a focus on feature importance analysis and fairness evaluation to ensure ethical AI use in healthcare.*

3. Flowchart of the Project Workflow

The detailed workflow is represented in the flowchart format to specify the each step taken in this project work



4. Data Description

- *Dataset Name & Origin: The project uses publicly available datasets such as the UCI Heart Disease Dataset, Kaggle Diabetes Dataset, and API-based real-time patient data (optional).*
- *Type of Data: Structured tabular data containing numeric and categorical patient attributes (e.g., age, blood pressure, glucose level).*
- *Number of Records and Features:*
  - *UCI Heart Disease Dataset has ~300 records with ~14 features.*
  - *Kaggle Diabetes Dataset includes ~750 records and 9 features.*
  - *Static or Dynamic: Primarily static datasets, though integration with dynamic real-time data APIs is possible for future work.*
  - *Target Variable: Presence or absence of a particular disease (binary classification), such as:*
    - *Diabetes: Outcome*
    - *Heart Disease: target*

5. Data Preprocessing

We performed essential data cleaning and preparation steps to ensure high-quality input for our predictive models.

- *Missing Values: Imputed missing values using mean/mode strategies for clinical metrics; dropped records with excessive missing data.*
- *Duplicates: Identified and removed duplicate patient records to prevent data leakage.*
- *Outliers: Detected and treated abnormal values in features like age, blood pressure, and glucose levels using IQR and Z-score methods.*
- *Data Types: Standardized data types (e.g., converting age to integer, date formats).*
- *Categorical Encoding: Applied label encoding and one-hot encoding to features like gender, symptoms, and diagnosis history.*
- *Normalization: Scaled features using MinMaxScaler to bring all values to a similar range.*
- *Documentation: Each transformation was clearly recorded with justification.*

6. Exploratory Data Analysis (EDA)

We conducted comprehensive statistical and visual analysis to uncover patterns in patient data.

- *Univariate Analysis: Explored distributions of individual features using histograms, boxplots, and countplots.*
- *Bivariate/Multivariate Analysis: Used correlation matrix, scatterplots, and pair plots to identify relationships between features and with the target disease outcome.*
- *Insights Summary: Highlighted key trends such as common symptoms for certain diseases, strong correlations between vitals and disease occurrence, and influential features for prediction.*

7. Feature Engineering

We enhanced the dataset to improve model accuracy and relevance.

- *Created new features such as BMI from height and weight, and risk score from combined vitals.*
- *Combined features like date of diagnosis with age to derive time-based features.*
- *Applied binning to age and glucose levels for categorical analysis.*
- *Used PCA to reduce dimensionality and improve model efficiency.*
- *Justified each feature addition/removal based on domain knowledge and correlation with the target variable.*

8. Model Building

To address the disease prediction problem, we implemented two machine learning models: Random Forest and Logistic Regression.

Model Choice Justification:

- *Random Forest was chosen for its robustness, ability to handle feature importance, and suitability for imbalanced healthcare data.*

- *Logistic Regression offers interpretability and performs well on binary classification tasks like disease/no disease.*

- *Data was split into training (80%) and testing (20%) sets with stratification to maintain class distribution.*
  Models were trained and evaluated using classification metrics: accuracy, precision, recall, and F1-score.

9. Visualization of Results & Model Insights

Visual tools such as confusion matrix, ROC curves, and feature importance plots were used to interpret model behavior.

Comparative plots illustrated Random Forest outperforming Logistic Regression in recall and F1-score.
Top features influencing predictions included age, blood pressure, glucose level, and cholesterol.
Each plot was clearly annotated to support conclusions about model reliability and feature impact.

10. Tools and Technologies Used

- *Programming Language: Python*
- *IDE/Notebooks: Google Colab, Jupyter Notebook*
- *Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, XGBoost*
- *Visualization Tools: Plotly, Tableau*

11. Team Members and Contributions

- *Data Cleaning: [Priyadharshini R] – Removed missing values, handled outliers, standardized formats.*
- *EDA: [Tejashwini P R] – Explored distributions, correlations, and health indicators.*
- *Feature Engineering: [Swetha K] – Created derived features from patient vitals and history.*
- *Model Development: [Preethi R] – Built and optimized machine learning models.*
- *Documentation and Reporting: [Sandhiya S] – Prepared final reports, plots, and insights summary.*