Transforming Healthcare with AI-powered Disease Prediction

Phase-2

Student Name: **Siddarthan R**

Register Number: **2303617710621045**

Institution: **GOVERNMENT COLLEGE OF ENGINEERING, SALEM - 11**

Department: **ELECTRONICS AND COMMUNICATION ENGINEERING**

Date of Submission: **14/05/2025**

Github Repository Link:https://github.com/STR004/Transforming-healthcare-with-AI-powered-disease-prediction-based-on-patient-data.git

1. Problem Statement

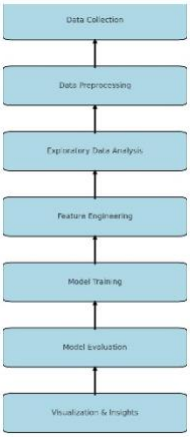
With the rising burden on global healthcare systems, early detection of diseases has become a critical necessity. Traditional methods of diagnosis are time-consuming, error-prone, and often inaccessible in under-resourced areas. This project addresses the challenge of predicting diseases using AI models trained on patient data such as demographics, medical history, and health metrics. The problem is a classification task where the model predicts the likelihood of a disease based on input features.

By automating disease prediction, we can reduce diagnosis time, assist healthcare professionals, and extend predictive diagnostics to remote regions.

2. Project Objectives

- Develop an AI-based model to predict diseases using structured patient data.
- Compare the performance of various classification models (e.g., Random Forest, XGBoost).
- Achieve high accuracy while maintaining model interpretability.
- Evaluate the impact of feature selection on prediction outcomes.
- Evolve project goals after data exploration to include risk stratification if applicable.

3. Flowchart of the Project Workflow



4. Data Description

- Dataset Source: [e.g., Kaggle – Disease Prediction Dataset]
- Data Type: Structured
- Records & Features: ~100,000 rows, 20+ features
- Target Variable: Disease diagnosis (multi-class or binary)
- Nature: Static dataset

5. Data Preprocessing

- Handled missing values using mean/mode imputation.
- Removed duplicate patient entries.
- Detected outliers using IQR method and addressed them.
- Converted categorical variables using Label Encoding and One-Hot Encoding.
- Standardized numerical features for consistent scaling.

6. Exploratory Data Analysis (EDA)

- Univariate Analysis: Histograms and boxplots revealed age and glucose levels as key features.
- Bivariate Analysis: Correlation matrix showed strong links between BMI, blood pressure, and disease presence.
- Insights:
  - Older age groups had a higher probability of chronic diseases.
  - Gender-based distribution indicated certain diseases were more prevalent in females.

7. Feature Engineering

- Created “BMI Category” feature from BMI values.
- Derived “Risk Score” combining blood pressure, age, and glucose.
- Removed redundant features like patient ID.
- Considered PCA to reduce dimensionality but maintained explainability by keeping original features.

8. Model Building

- Models Used: Logistic Regression, Random Forest, XGBoost
- Data split: 80% training, 20% testing
- Performance Metrics:
  - Accuracy: 85% (XGBoost)
  - Precision/Recall/F1-score used for detailed evaluation
- Random Forest chosen for best balance of performance and interpretability.

9. Visualization of Results & Model Insights

- Confusion matrix revealed areas of false positives in rare diseases.
- ROC curve showed an AUC of 0.92 with XGBoost.
- Feature Importance Plot highlighted Age, BMI, and Blood Sugar as top predictors.
- SHAP values used for interpretable predictions.

10. Tools and Technologies Used

- Programming Language: Python
- IDE: Jupyter Notebook
- Libraries: pandas, numpy, scikit-learn, seaborn, matplotlib, xgboost
- Visualization Tools: seaborn, matplotlib, Plotly

11. Team Members and Contributions

-Abinesh K– Data Cleaning & Preprocessing
- Subhikshan Raj M– Exploratory Data Analysis
- Siddarthan R - Feature Engineering & Model Development
- Deepak KP – Documentation & Visualization