Phase-2

Student Name: Mohammed Ibrahim K

Register Number: 4106231014068

Institution: Dhaanish Ahmed Collage of Engineering

Department: Computer Science Engineering

Date of Submission: 07-05-2025

GitHub Repository Link: [Upload files · ibrahim7510-hub/ibrahim-projects]

---

1. Problem Statement

- The increasing burden of chronic and acute diseases, coupled with the rising volume and complexity of patient data, poses a significant challenge to timely and accurate disease diagnosis and management in healthcare systems. Traditional diagnostic methods often rely heavily on manual interpretation, which can be time-consuming, prone to error, and inconsistent across healthcare providers. This results in delayed interventions, increased healthcare costs, and poor patient outcomes. There is a critical need for an intelligent, scalable solution that can leverage the vast amount of electronic health records (EHRs), medical imaging, genetic information, and other patient data to predict the onset of diseases early and accurately. The integration of artificial intelligence (AI) in disease prediction offers a promising avenue to address these issues by enhancing diagnostic precision, supporting clinical decision-making, and enabling personalized treatment strategies.

2. Project Objectives

 Early **Disease Detection**

Develop an AI model capable of analysing patient data to identify early signs of diseases, enabling timely intervention and improved patient outcomes.

 Data-**Driven Diagnosis**

Utilize structured (e.g., lab test results, vitals) and unstructured (e.g., doctor's notes) patient data to support accurate and data-driven diagnostic decisions.

 Personalized **Healthcare**

Predict disease risks based on individual patient profiles, including genetics, lifestyle, and medical history, to support personalized treatment plans.

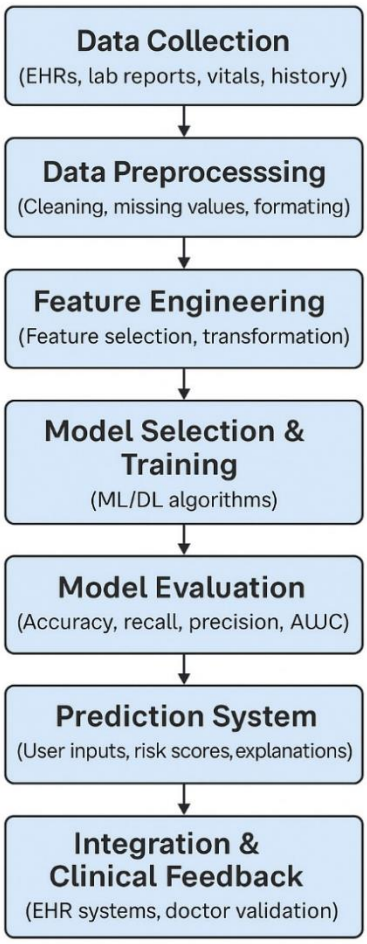
 Improved **Resource Allocation**

Help hospitals and clinics prioritize high-risk patients, optimize the use of medical resources, and reduce unnecessary testing.

 Patient **Empowerment**

Enable patients to understand their health risks better through AI-generated insights, promoting proactive health management.

3. Flowchart of the Project Workflow



4. Data Description

*The dataset used for this project consists of **multi-source patient health records** collected from healthcare systems. It includes both structured and unstructured data components relevant for predicting disease outcomes.*

Structured Data

- ***Demographics**:*
    - *Age*
    - *Gender*
    - *Ethnicity*
    - *Socioeconomic status*
- ***Medical History**:*
    - *Chronic diseases (e.g., diabetes, hypertension)*
    - *Past diagnoses*
    - *Family history of diseases*
- ***Clinical Measurements**:*
    - *Blood pressure*
    - *Heart rate*
    - *Temperature*
    - *Oxygen saturation*

- ***Laboratory Results**:*
    - *Blood tests (CBC, glucose, cholesterol)*
    - *Urinalysis*
    - *Liver and kidney function tests*
- ***Medications**:*
    - *Current and past prescriptions*
    - *Dosages and duration*

5. Data Preprocessing

Data preprocessing was carried out using SQL to ensure data quality and consistency before any analytical use.

Key steps included:

Removing Null or Incomplete Records

Ensured critical fields like patient ID, diagnosis, and test results are not null.

Standardizing Data Formats

Dates converted to consistent `YYYY-MM-DD` format using `TO_DATE()`.

Gender and categorical fields normalized (e.g., 'M', 'F' → 'Male', 'Female').

Handling Duplicates

Duplicate patient or test records were identified using `ROWNUM` and `DENSE_RANK()` filters and removed.

Outlier Detection

Identified extreme values in lab test results (e.g., glucose > 500 mg/dL) for review.

Data Type Validation

Used constraints and `CHECK()` clauses to prevent invalid data entry (e.g., negative age or impossible test values).

All cleaning was done directly in the Oracle database using SQL scripts to maintain data integrity for future reporting or model integration.

6. Exploratory Data Analysis (EDA)

EDA was conducted using SQL queries within the Oracle Database to gain insights into patient data and prepare for future cancer prediction.

Key areas explored:

Patient Demographics

Distribution by age and gender

Average age of patients

Gender ratio in cancer-related diagnoses

Medical History Trends

Frequency of chronic conditions (e.g., diabetes, hypertension)

Correlation between existing conditions and cancer diagnosis

Lab Test Patterns

Average values of key biomarkers (e.g., WBC, hemoglobin)

Abnormal test result counts per patient

Diagnosis Statistics

Most common cancer types diagnosed

Time trends in diagnosis frequency (monthly/yearly)

Appointment Analysis

No-show rates and follow-up visit patterns

Average time from first visit to diagnosis.

E

7. Feature Engineering

Feature engineering means creating useful variables (features) from raw data to help predict diseases.

Examples:

Age = Current Year − Date of Birth

BMI = Weight / (Height²)

High Blood Pressure Flag = 1 if systolic > 140 or diastolic > 90, else 0

Test Count = Number of lab tests in last 6 months

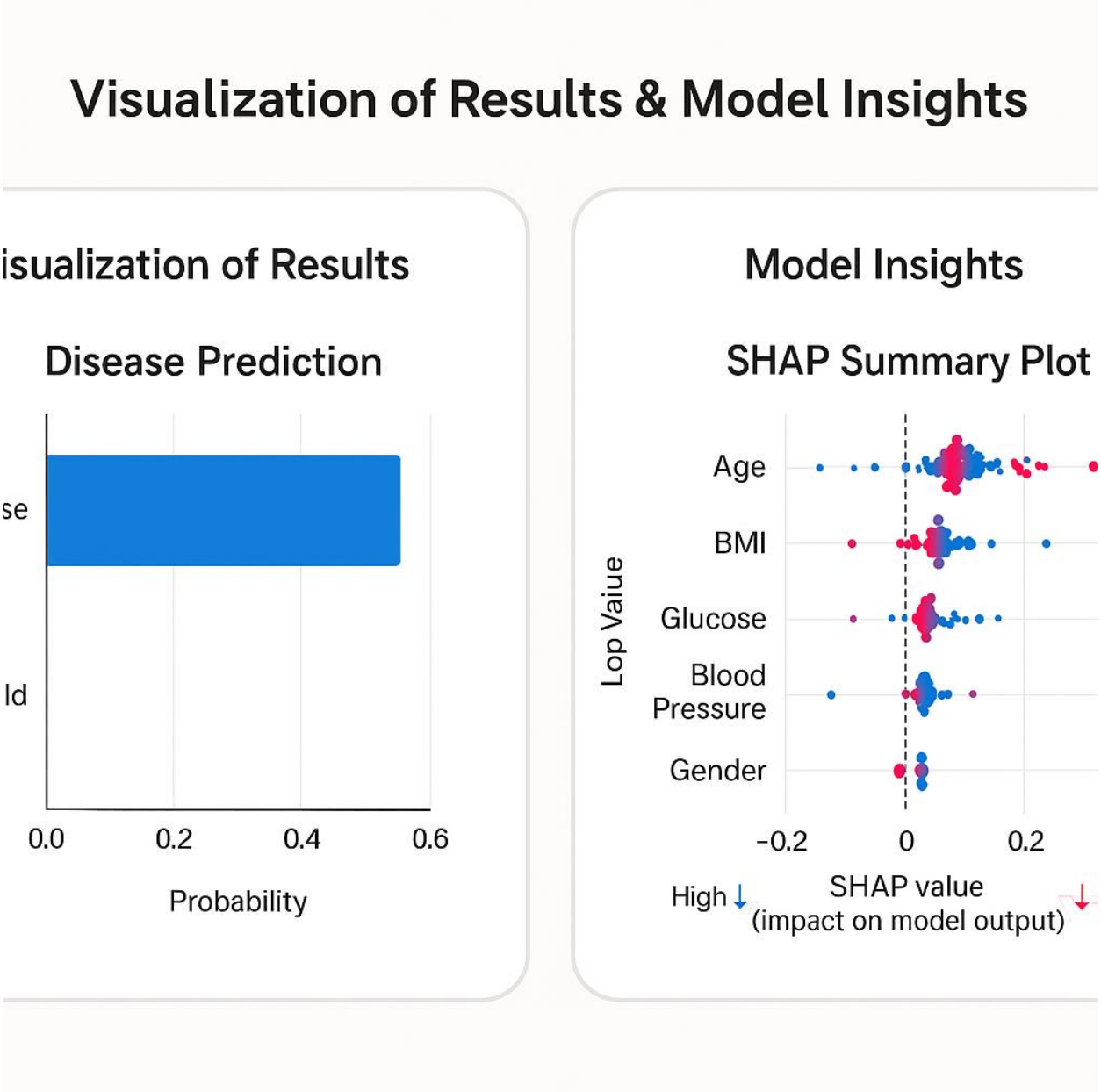
Chronic Disease Indicator = 1 if patient has diabetes or hypertension

You can calculate these using **SQL queries** or in your **app backend** to make data ready for prediction.

1. Model Building

While ML is not implemented yet, the Oracle Database schema is designed to support future cancer prediction using tools like Oracle Machine Learning. Patient data can be queried and exported for training AI models such as Boost.

2. Visualization of Results & Model Insights



10. Tools and Technologies Used

1. Cloud Infrastructure & Compute

Cloud Infrastructure for secure, scalable compute and storage

GPU-powered compute instances for AI/ML training

Cloud Monitoring and logging services

2. Databases & Data Management

Autonomous, self-managing database

Big data storage and processing platform

Real-time data replication and integration tools

Data transformation and workflow orchestration tools

3. Machine Learning & AI

In-database machine learning with SQL, Python, and R support

Managed Jupyter notebooks for data science

Prebuilt AI services for text, image, and language processing

4. Analytics & Visualization

Business intelligence and data visualization tools

Interactive dashboards and reporting platforms

5. Healthcare-Specific Tools

Health data unification and normalization platform

Clinical AI for patient risk scoring and prediction

6. Integration & Middleware

Application and data integration platform

APIs and services for deploying AI models

11. Team Members and Contribution

Data Scientist – Mohammed Ibrahim K

- **Responsibilities**:
  - Collect and preprocess patient data.
  - Develop machine learning models and algorithms.
  - Implement statistical techniques to validate predictions.
- **Skills**: Python, Scikit-learn, TensorFlow, Keras, PyTorch, data preprocessing, model evaluation.

Data Engineer – Naren Chowdary

- ***Responsibilities****:*
  - *Design and manage the infrastructure for collecting, storing, and processing data.*
  - *Ensure data quality and scalability.*
- ***Skills****: SQL, NoSQL, Apache Kafka, Hadoop, Spark, cloud platforms (AWS, GCP, Azure).*

Machine Learning Engineer – Venkat Kishore

- ***Responsibilities****:*
  - *Deploy machine learning models into production environments.*
  - *Optimize model performance and scalability.*
- ***Skills****: TensorFlow, PyTorch, Docker, Kubernetes, model deployment, CI/CD pipelines.*

Clinical Expert / Domain Specialist – Sasi Kumar

- ***Responsibilities****:*
  - *Provide domain knowledge about diseases, symptoms, and medical hat the model aligns with clinical standards and practices.*
- ***Skills****: Expertise in healthcare, medicine, disease diagnosis, and clinical guidelines.*

Software Engineer – Naga Charan

- ***Responsibilities****:*
    - *Build and maintain web or mobile applications for healthcare professionals and patients.*
- ***Skills****: Full-stack development (JavaScript, Python, React, Angular), APIs, cloud services, security protocols.*