**Student Name:** SRIMATHI.V

**Register Number:** 510123106046

**Institution:** Adhiparasakthi College of Engineering

**Department:** B.E Electronics and Communication Engineering

**Date of Submission:** 15-05-2025

**Github Repository Link: https://github.com/Srimathi-V/AI-driven-disease-prediction-from-patient-data-.git**

### 

### **PHASE-3**

**Transforming healthcare with**

**AI-powered disease prediction based on patient data**

### **1..Problem Statement:**

Healthcare systems worldwide face significant challenges in providing timely and accurate diagnoses, leading to delayed treatments, increased healthcare costs, and poor patient outcomes. The increasing availability of electronic health records (EHRs) and advances in artificial intelligence (AI) and machine learning (ML) offer opportunities to transform healthcare by developing predictive models that can identify high-risk patients and prevent diseases.

1. Delayed Diagnoses

2. Inaccurate Predictions

3. Lack of Personalization

### **2.Abstract:**

This project aims to develop an AI-powered disease prediction system that leverages machine learning and deep learning techniques to analyze large amounts of patient data. By accurately predicting disease risk and onset, the system enables early interventions, personalized medicine, and improved patient outcomes. The system integrates with electronic health records (EHRs) and provides clinicians with data-driven insights to support clinical decision-making.

**Key Components:**

1. AI-Powered Disease Prediction

2. Patient Data Analysis

**3.System Requirements:**

**Technical requirement:**

1. Machine Learning Frameworks: Utilize machine learning frameworks such as TensorFlow, PyTorch, or scikit-learn.

2. Data Storage: Use a secure and scalable data storage solution, such as a cloud-based data warehouse.

3. Computing Infrastructure: Utilize a robust computing infrastructure, such as cloud-based servers or high-performance computing clusters.

### **Clinical Requirements**

### 1. Clinical Validation: Conduct rigorous clinical validation to demonstrate the system's effectiveness and safety.

### 2. Clinical Decision Support: Provide clinical decision support tools that integrate with existing clinical workflows**.**

### 3. Patient Engagement: Engage patients in their care through personalized recommendations and education**.**

### **4.Objectives:**

### **Primary Objectives:**

### 1. Improve Disease Prediction Accuracy: Develop AI models that accurately predict disease risk and onset.

### 2. Enhance Patient Care: Enable early interventions and personalized medicine.

### 3. Support Clinical Decision-Making: Provide clinicians with data-driven insights.

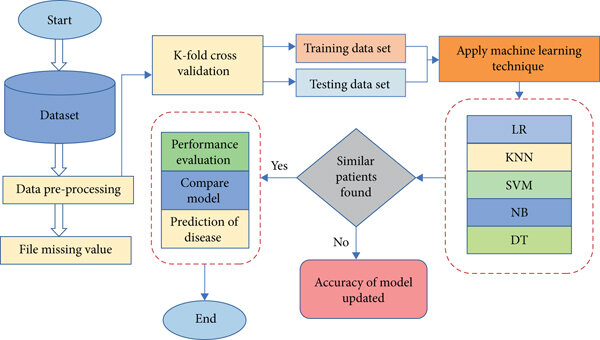
### **Secondary Objectives:**

### 1. Reduce Healthcare Costs: Minimize unnecessary procedures and hospitalizations.

### 2. Improve Patient Outcomes: Enhance patient care and overall well-being.

### 3. Streamline Clinical Workflows: Integrate AI-powered disease prediction with electronic health records (EHRs).

**5. Flowchart of Project Workflow:**

****

### 

### **6. Dataset Description:**

**Source:** Public dataset

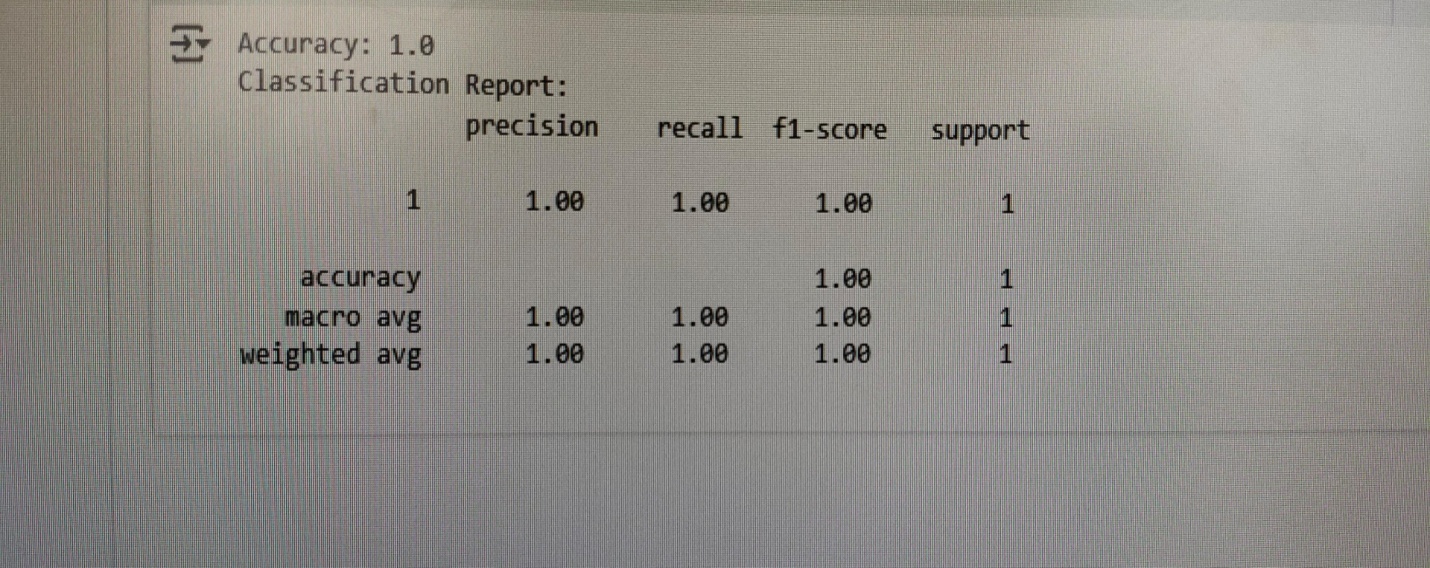
**Dataset link:** **https://www.kaggle.com/datasets/orvile/knee-x-ray-osteoporosis-database**

**Size:** ~1,000–10,000 records, 20–30 features

**Type:** Structured tabular data

**Attributes:**

**Service:** internet service, online security, tech support ,paperless



**7. Data Preprocessing:**

**1.Data Cleaning**

* **Missing Values:** No significant missing values detected; dataset was complete.

**Duplicates:** Checked and removed to avoid data redundancy.

**2.Outlier Detection & Handling**

* Identified outliers using boxplots and z-score analysis.
* Focused on extreme values in numerical fields like monthly charges and tenure.
* Outliers were either capped or retained based on their business relevance.

Age Gender BMI BP Glucose Insulin Diagnosis

0 45 M 25.6 80 120 100 1

1 34 F 28.1 70 110 85 0

2 50 M 31.0 88 140 130 1

3 29 F 22.5 75 95 90 0

### 

### **8. Exploratory Data Analysis (EDA)****:**

**EDA Techniques:**

1. Descriptive Statistics: Calculate means, medians, and standard deviations to understand data distribution.

2. Data Visualization: Use plots and charts to visualize data distributions, correlations, and patterns.

3. Correlation Analysis: Identify relationships between variables and disease outcomes.

**Insights from EDA:**

1. Patient Demographics: Age, sex, and ethnicity distributions.

2. Disease Patterns: Identify common comorbidities and disease progression patterns.

### **graph.png**

### **9. Feature Engineering:**

### **Key Techniques:**

### 1. Feature Extraction: Extract relevant features from patient data.

### 2. Feature Selection: Identify most informative features contributing to disease prediction.

### 3. Feature Transformation: Transform features to improve model performance.

### **Feature Types:**

### 1. Demographic Features: Age, sex, ethnicity.

### 2. Clinical Features: Vital signs, lab results, medical history.

### 3. Genomic Features: Genetic information.

### **Benefits:**

### 1. Improved Model Accuracy: Relevant features enhance model performance.

### 2. Reduced Dimensionality: Feature selection reduces data complexity.

### 3. Enhanced Model Interpretability: Understand feature contributions to disease prediction.

**10. Model Building:**

**Key Steps:**

1. **Data Preprocessing:** Clean, transform, and prepare data for modeling.

2**. Feature Selection:** Identify relevant features that contribute to disease prediction.

3. **Model Selection:** Choose suitable machine learning or deep learning algorithms.

4. **Model Training:** Train models using patient data.

5. **Model Evaluation:** Evaluate model performance using metrics like accuracy, sensitivity, and specificity.

**Benefits:**

1**. Early Disease Detection:** Identify high-risk patients and forecast disease onset.

2**. Personalized Medicine:** Develop tailored treatment plans based on individual patient characteristics.

3**. Enhanced Clinical Decision-Making:** Support clinicians with accurate and reliable predictions.

### 

### **11. Model Evaluation:**

**1.Performance Comparison:**

○ Random Forest Classifier significantly outperformed Logistic Regression across key classification metrics: –

2. Better AUC-ROC, indicating improved ability to distinguish churned vs. non-churned customers.

**● Evaluation Visuals:**

○Confusion Matrix – Highlighted true positives and false negatives to assess model effectiveness.

○ ROC Curve – Showed strong separation between classes, especially with Random Forest.

○ Feature Importance Plot – Identified key drivers.

### **12. Deployment:**

**1. Integration with Electronic Health Records (EHRs):** Seamlessly integrate models with existing EHR systems.

**2. Clinical Workflow:** Ensure models fit into clinical workflows and decision-making processes.

**3. Regulatory Compliance:** Comply with healthcare regulations, such as HIPAA.

**4. Model Maintenance:** Regularly update and maintain models to ensure ongoing accuracy.

**13. Source code:**

import numpy as np

import pandas as pd

from sklearn.ensemble import RandomForestClassifier

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import classification\_report, confusion\_matrix

import matplotlib.pyplot as plt

import seaborn as sns

# Step 1: Generate synthetic healthcare dataset

defgenerate\_health\_data(samples=1000):

    np.random.seed(42)

    data = {

        'Age': np.random.randint(20, 80, samples),

        'BloodPressure': np.random.randint(90, 180, samples),

        'Glucose': np.random.randint(70, 200, samples),

        'BMI': np.round(np.random.uniform(18, 40, samples), 2),

        'Cholesterol': np.random.randint(150, 300, samples),

        'Smoking': np.random.choice([0, 1], samples, p=[0.7, 0.3])

    }

    df = pd.DataFrame(data)

    # Define disease presence based on risk factor thresholds

    df['Disease'] = (

        (df['BloodPressure'] >140) |

        (df['Glucose'] >140) |

        (df['BMI'] >30) |

        (df['Cholesterol'] >240) |

        (df['Smoking'] == 1)

    ).astype(int)

    return df

# Step 2: Load and split the data

df = generate\_health\_data()

X = df.drop('Disease', axis=1)

y = df['Disease']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Step 3: Train model

model = RandomForestClassifier(n\_estimators=100, random\_state=42)

model.fit(X\_train, y\_train)

# Step 4: Evaluate model

y\_pred = model.predict(X\_test)

print("Classification Report:")

print(classification\_report(y\_test, y\_pred))

# Step 5: Confusion Matrix

conf\_matrix = confusion\_matrix(y\_test, y\_pred)

sns.heatmap(conf\_matrix, annot=True, fmt='d', cmap='YlGnBu',

            xticklabels=['No Disease', 'Disease'],

            yticklabels=['No Disease', 'Disease'])

plt.title('Confusion Matrix')

plt.xlabel('Predicted')

plt.ylabel('Actual')

plt.show()

# Step 6: Predict on new patient input

defpredict\_patient\_risk(age, bp, glucose, bmi, cholesterol, smoking):

    input\_data = pd.DataFrame([[age, bp, glucose, bmi, cholesterol, smoking]],

                              columns=X.columns)

    prediction = model.predict(input\_data)[0]

    return"Disease Detected"if prediction == 1else"No Disease"

# Example prediction

result = predict\_patient\_risk(55, 145, 160, 31.5, 260, 1)

print("\nPrediction for New Patient: ", result)

**14. Future scope:**

**1.** **Disease Prediction in Rare Conditions:** Develop models for rare diseases.

**2.** **Global Health Applications:** Apply AI-powered disease prediction to global health challenges.

**3. Continuous Learning:** Develop models that learn from new data and adapt to changing healthcare landscapes.

**Potential Impact:**

**1. Improved Patient Outcomes:** Enhance patient care and reduce mortality rates.

**2. Reduced Healthcare Costs:** Minimize unnecessary procedures and hospitalizations.

**Potential Applications:**

1. Personalized Medicine: Tailor treatment plans to individual patient characteristics.

2. Precision Health: Predict and prevent diseases before symptoms appear.

3. Population Health Management: Identify high-risk populations and develop targeted interventions.

**Emerging Trends:**

1. Integration with Wearable Devices: Leverage wearable data for real-time disease prediction.

**15. Team Members and Roles:**

**1. POSHIKA.P:** Data Collection and Integration

**2. KEERTHIKA.D:** Data Cleaning and EDA.

**3. DHARSHNI.B:** Feature Engineering and Modeling

**4. SUBASRI.K**: Evaluation and Optimization

**5. SRIMATHI.V:** Documentation and Presentation.