Machine Learning

ABSTRACT

Predictive Disease Diagnosis Using Machine Learning

Team Members

Name	Roll No
K. Sai Dinesh	AM.SC.U4CSE23130
Satya Sri Dheeraj Motupalli	AM.SC.U4CSE23150
Y. Sai Nikhil	AM.SC.U4CSE23171
P. Sai Satvik	AM.SC.U4CSE23142

Introduction

Early diagnosis of diseases plays a crucial role in improving treatment outcomes and reducing health-care costs. Traditional diagnostic methods often rely on manual interpretation and can be time-consuming or prone to human error. With the increasing availability of medical data, machine learning provides an efficient way to identify disease patterns and make accurate predictions. This project focuses on applying machine learning algorithms to predict the likelihood of various diseases-such as Heart Disease, Breast Cancer, Diabetes, and AIDS-based on relevant medical features and patient data.

Problem Statement

Despite advancements in health-care, timely detection of diseases remains a challenge due to the complexity and volume of patient data. Medical professionals often lack tools that can assist in quick and accurate prediction. The problem addressed in this project is the development of machine learning models capable of predicting the presence or risk of common diseases using medical datasets. Each disease requires unique data preprocessing, feature selection, and model tuning to ensure high diagnostic accuracy.

Project Objective

1. To collect and preprocess disease-specific medical datasets (Heart Disease, Breast Cancer, Diabetes, AIDS).
2. To apply and compare the performance of 4–5 suitable machine learning algorithms for each disease.
3. To evaluate model performance using metrics like accuracy, precision, recall, and F1-score.
4. To analyze the key medical features influencing each disease outcome.
5. To demonstrate how ML can support healthcare professionals in early and data-driven diagnosis.

Dataset and Member Allocation

1. AIDS Virus Infection Prediction [K. Sai Dinesh]

- **Dataset:** AIDS Virus Prediction Dataset
- **Description:** Includes clinical and demographic features like age, blood parameters, viral count, and behavioral indicators related to HIV/AIDS infection.
- **Planned Algorithms:**
  Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Gradient Boosting.
- **Objective:** To classify whether a patient is infected with the AIDS virus and evaluate algorithmic robustness on medical data.

2. Diabetes Prediction [Satya Sri Dheeraj Motupalli]

- **Dataset:**Diabetes Prediction Dataset
- **Description:** Comprises physiological and lifestyle features such as glucose level, BMI, age, insulin, and family history of diabetes.
  Planned Algorithms:
  Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Naïve Bayes.
- **Objective:** To predict whether an individual is likely to develop diabetes based on medical and lifestyle parameters.

3. Heart Disease Prediction [Y. Sai Nikhil]

- **Dataset:** Heart Disease Dataset
- **Description:** Contains patient attributes such as age, cholesterol, blood pressure, ECG readings, and exercise-induced angina.
- **Planned Algorithms:**
  Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting.
- **Objective:** To predict the presence of heart disease based on patient clinical data and determine which algorithm offers the highest diagnostic accuracy.

4. Breast Cancer Detection [P. Sai Satvik]

- **Dataset:** Breast Cancer Dataset
- **Description:** Features extracted from breast tissue samples such as radius, texture, smoothness, and compactness to classify tumors as benign or malignant.
  Planned Algorithms:
  Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Naïve Bayes.
- **Objective:** To classify tumors as malignant or benign and compare algorithmic accuracy and interpretability.

Expected Outcomes

- The study is expected to identify the most accurate and reliable machine learning algorithms for predicting disease occurrence based on patient health data.
- Each member will compare the performance of 4–5 models on their respective datasets.
- The combined results will highlight which algorithms perform best across different disease types in terms of accuracy, precision, recall, F1-score, and ROC-AUC.

Significance and Impact

This project emphasizes the growing role of machine learning in supporting early disease detection and medical diagnosis.
By applying a unified analytical framework to multiple health conditions, the study highlights how data characteristics affect model performance and predictive reliability.

The insights obtained can assist healthcare professionals in identifying high-risk patients and designing more efficient diagnostic tools.
Moreover, the project demonstrates how accessible machine learning approaches can enhance data-driven decision-making and promote the development of intelligent healthcare systems.