

AI-Based Multi Disease Prediction System Using Machine Learning

Divyalakshmi K, Student, Department of Computer Science and Engineering,

City Engineering College, divyalakshi21@gmail.com

K. Archana Bhat, Assistant Professor, Department of Computer Science and Engineering,

City Engineering College, archanabhat247@gmail.com

Tejaswini B N, Assistant Professor, Department of Computer Science and Engineering,

City Engineering College, tejaswini.begur@gmail.com

Abstract - Early diagnosis of chronic diseases is crucial in reducing mortality rates and enhancing patient outcomes, yet access to timely diagnostic screening remains limited in many settings. This framework presents an AI-based Multi-Disease Prediction that uses machine learning to assess the six major diseases Diabetes, Lung Cancer, Stroke, Parkinson's Disease, Heart Disease and Liver Disease through a single, unified platform.

The framework is built using Python and Streamlit, providing an interactive web-based framework an interface for entering clinical variables and receive real-time predictions. Each disease-specific model was trained independently on publicly available datasets (Kaggle, UCI ML Repository) using framework such as Random Forest selected based on comparative performance evaluation using metrics accuracy. the system incorporates an Explainable AI component that highlights the influencing each prediction, alongside a dedicated Risk Assessment that translates model outputs into interpretable risk levels.

This application supports automated PDF medical report generation via ReportLab and maintains a persistent patient record database using SQLite, enabling historical tracking of predictions. By integrating predictive modelling, explainability, and usability into one accessible tool, this system aims to support early risk identification and assist both patients and healthcare practitioners in making informed, data-driven decisions

serving as a step toward affordable and scalable preventive healthcare solutions.

1. INTRODUCTION

Access to timely and affordable diagnostic screening remains a persistent difficulty for healthcare systems around the world, the chronic and life-threatening conditions like diabetes, stroke, liver disease, lung cancer heart disease and Parkinson's. Preliminary detection of these illnesses generally leads to far better treatment outcomes, yet identifying them at an early stage typically depends on specialist medical knowledge, costly testing, and lengthy diagnostic workflows. Such resources are frequently out of reach in rural and underserved areas, leaving a huge portion of the population unaware that they may already be at risk until their condition has advanced considerably.

Recent progress in AIML has opened up new possibilities for identifying this gap. By training models on medical datasets, it becomes possible to uncover statistical relationships between a patient's symptoms, clinical parameters, and eventual disease outcomes. These models can be used to detect risk early, offering a practical way to extend preventive diagnosis to populations who currently lack easy access to specialized screening.

Building on this idea, the current project presents an AI-Based Multi-Disease Prediction System: a single web-based platform designed to estimate the six major diseases at once. Rather than relying on separate tools for

each condition the approach taken by most existing systems this project brings multiple prediction models together under one interface, making convenient for both patients and healthcare providers while cutting down on the need for multiple standalone tools.

The application itself is built in Python, using Streamlit for the front-end interface and Scikit-learn to power the prediction models. Beyond generating predictions, the system places particular emphasis on explainability and ease of use two qualities that are basically neglected in comparable healthcare ML tools. An Explainable AI module is integrated insight into the reasoning behind each prediction, thereby 18 improving user trust and transparency. Furthermore, a risk assessment module translates raw prediction outputs into interpretable risk categories, making the results more meaningful for non-technical users.

2. LITERATURE SURVEY

Machine learning framework has become an important technology in healthcare for disease diagnosis and predictive analysis. Several researchers have proposed intelligent systems that assist clinical professionals in identifying diseases at an early stage using clinical data.

1. Breiman (2001) introduced Random Forest method integrates numerous decision trees to identify overall predictive performance and minimize overfitting issues. Because of robustness and accuracy, Random Forest is used extensively in healthcare prediction model.
2. Kavakiotis et al. (2017) studied the model of Data-driven learning and mining methodologies for diabetes prediction. Their research demonstrated that predictive analytics can assist in identifying disease susceptibility factors and support early-stage

diagnosis.

3. Detrano et al. (1989) proposed a Machine Learning-Based framework for heart disease screening using the Cleveland data insights. This model revealed model framework can effectively analyse.
4. Boukhatem et al. (2022) proposed a cardiovascular prediction model that employed several classification techniques to identify heart-related conditions. The study of machine learning-based approaches can enhance the accuracy of prediction and tool for clinical decision-making and patient assessment.
5. Bhatt et al. (2023) conducted a comparative evaluation of frameworks for diagnosis heart disease. Their findings revealed that ensemble-based methods, particularly Random Forest, achieved superior predictive performance and accurate than traditional classification algorithms
6. Srivastava et al. (2025) proposed a multi-disease prediction model using Support Vector Machine (SVM), Random Forest and K-Nearest Neighbors (KNN). Their study highlighted that effective data preprocessing and feature selection significantly improve prediction performance across multiple diseases.

3. RESEARCH GAP

Machine learning has use in disease prediction, majority of healthcare systems currently in use are built to identify just one condition at a time diabetes or heart disease, for instance which restricts how useful they are in real clinical settings, where several health issues often need to be evaluated together. Because most prediction models are built as standalone applications for a single disease, users end up juggling multiple tools, which adds

unnecessary complexity and cuts down on overall efficiency.

In most cases, these systems simply output a prediction without revealing which factors drove that outcome. This opacity makes it harder for both clinicians and patients to trust AI-generated diagnostic results. While techniques like SHAP and LIME have emerged as ways to make Explainable AI more accessible, they are still rarely built into real-world healthcare prediction tools.

Beyond this, a lot of current applications stop at prediction and skip other functionality that a complete clinical decision-support system would need things like risk scoring, automatic report creation, or maintaining patient records. Much of the existing research prioritizes prediction accuracy alone, rather than building something end-to-end that is genuinely practical, scalable, and easy for users to work with.

This project puts forward an AI-Based Multi-Disease Prediction System together several disease-prediction models inside one Streamlit-based application. Along with Explainable AI, the system also incorporates risk assessment, PDF report generation, and an SQLite-based patient database, together forming a transparent, efficient, and user-friendly solution for preliminary disease detection and clinical decision support.

4. PROBLEM STATEMENT

The increasing prevalence of chronic diseases such as Diabetes, Heart Disease, Stroke, Lung Cancer, Liver Disease, and Parkinson's Disease has created a significant need for early and accurate diagnosis.

Most existing disease prediction systems are designed to predict only a single disease and lack features such as explainable predictions, risk assessment, automated report generation, and patient record management. As a

result, healthcare professionals and patients must rely on multiple independent applications, leading to increased complexity and reduced efficiency.

To address these challenges, there is a need for an integrated, intelligent, and user-friendly healthcare system capable of predicting multiple diseases through a single platform. The proposed AI-Based Multi-Disease Prediction System utilizes the Random Forest Classifier to provide real-time disease prediction, Explainable AI, risk assessment, PDF report generation, and patient database management, thereby supporting early disease detection and informed healthcare decision-making.

5. OBJECTIVES

The primary objective of the AI-Based Multi-Disease Prediction System is to develop an intelligent and user-friendly healthcare application that utilizes machine learning to predict the likelihood of multiple diseases at an early stage. The system aims to assist patients and healthcare professionals by providing fast, accurate, and reliable preliminary disease predictions.

The specific objectives of the project are:

1. To develop a unified healthcare platform capable of predicting multiple diseases, including Diabetes, Heart Disease, Stroke, Lung Cancer, Liver Disease, and Parkinson's Disease.
2. To implement the Random Forest Classifier for accurate and efficient disease prediction using patient medical data.
3. To provide real-time disease predictions through an interactive web application developed using Streamlit.
4. To perform data preprocessing and feature selection to improve the performance and accuracy of the machine learning models.

5. To incorporate Explainable AI (XAI) for identifying the key factors influencing prediction results, thereby improving transparency and user trust.
6. To assess the risk level associated with the predicted disease, enabling users to better understand their health condition.
7. To generate downloadable PDF medical reports containing patient details, prediction results, risk assessment, and prediction summary.
8. To maintain patient records and prediction history using an SQLite database for efficient data storage and retrieval.
9. To provide a simple, fast, and accessible healthcare decision-support system that can assist in early disease detection and promote preventive healthcare.
10. To create a scalable and extensible platform that can be enhanced in the future by integrating additional disease prediction models, cloud deployment, mobile applications, and hospital information systems.

6. METHODOLOGY

The proposed methodology consists of several sequential stages that transform the healthcare data into meaningful disease predictions.

Stage 1: Dataset Collection

Healthcare datasets were accumulated from publicly available repositories like the Kaggle and UCI Machine Learning Repository. Distinct datasets were identified for each disease to ensure disease-specific model training.

The datasets include:

- Diabetes Dataset
- Heart Disease Dataset
- Stroke Dataset
- Lung Cancer Dataset
- Liver Disease Dataset
- Parkinson's Disease Dataset

Each dataset contains clinical and demographic attributes relevant to disease diagnosis.

Stage 2: Data Preprocessing

Raw healthcare datasets may include missing values, duplicate records, inconsistent formats, and irrelevant features. Therefore, preprocessing is performed before model training.

The preprocessing steps include:

- Removing duplicate records
- Handling missing values
- Encoding categorical variables
- Data normalization (where required)
- Feature selection
- Train-test data splitting

Proper preprocessing improves model performance and reduces prediction errors.

Stage 3: Model Training

After preprocessing, separate Random Forest Classifier framework are trained for each disease dataset using the Scikit-learn library.

Random Forest is selected because it:

- Provides high prediction accuracy
- Reduces overfitting through ensemble learning
- Handles large healthcare datasets efficiently
- Supports feature importance analysis

- Performs well with nonlinear relationships

Each trained model is serialized using Joblib and stored inside the model/ directory for deployment.

Stage 4: Real-Time Disease Prediction

The trained models are implemented into a Streamlit application. The prediction process consists of the following steps:

1. User selects a disease.
2. User enters the required medical parameters.
3. Input values are validated.
4. Corresponding Random Forest model is loaded.
5. Prediction is generated instantly.
6. Prediction result is displayed.

The application predicts the likelihood of:

- Diabetes
- Stroke
- Heart Disease
- Lung Cancer
- Liver Disease
- Parkinson's Disease

Stage 5: Explainable AI

To enhance model transparency and user confidence, the system includes an Explainable AI (XAI) module. The Explainable AI module:

- Identifies important medical features
- Explains prediction outcomes
- Improves transparency
- Increases trust in machine learning predictions

This helps users to identify the variable influencing the model's decision rather than providing only a binary prediction.

Stage 6: Risk Assessment

The Risk Assessment Module evaluates the predicted disease outcome and categorizes the patient's condition into different risk levels.

The module provides additional insights regarding the severity of the predicted condition, assisting users in interpreting the prediction results more effectively.

Stage 7: PDF Report Generation

After prediction, the system automatically generates a downloadable PDF medical report containing:

- Patient information
- Disease name
- Prediction result
- Level of Risk
- Prediction summary
- Date and time of prediction

The reports are stored in the reports/ directory and can be downloaded for future reference.

Stage 8: Patient Database Management

Patient details and prediction history are securely stored using an SQLite database. The database stores:

- Patient Name
- Age
- Gender
- Disease Predicted
- Prediction Result
- Risk Level
- Date and Time

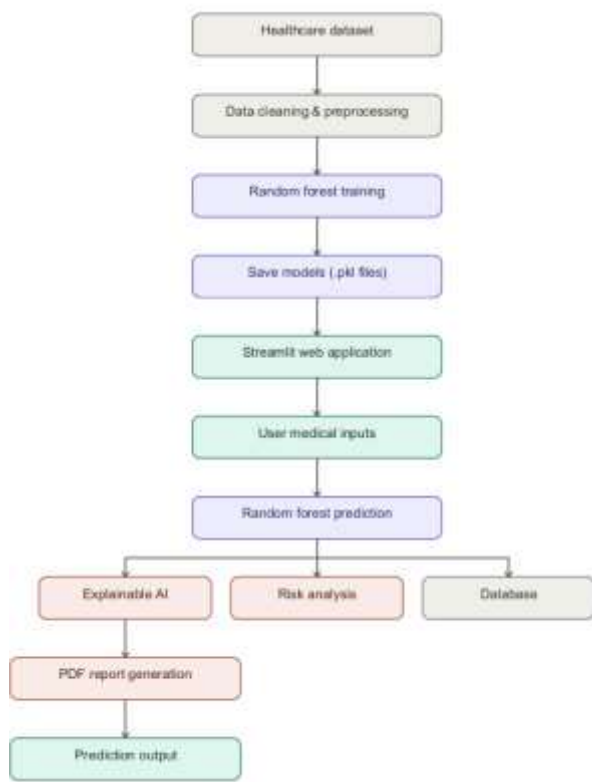
This enables efficient management and retrieval of previous prediction records.

Stage 9: System Deployment

The complete application is deployed using the Streamlit framework. The deployment process involves:

- Loading pre-trained Random Forest models
- Accepting user inputs
- Performing disease prediction
- Displaying Explainable AI results
- Generating PDF reports
- Updating the SQLite database

The web-based interface ensures accessibility, ease of use, and real-time prediction capabilities.



7. REQUIREMENT SPECIFICATION

The AI-Based Multi-Disease Prediction System was implemented using Python and several open-source libraries machine learning, data processing, web app development, database management, and report generation.

Python	The programming language used for developing the application and implementing machine learning algorithms.
Streamlit	Used to create an interactive and user-friendly web interface for disease prediction.
Sciklearn	Provides machine learning algorithms, especially the Random Forest Classifier
Pandas	Used for loading, cleaning, and manipulating healthcare datasets.
NumPy	Performs numerical computations and array operations required during data preprocessing.
Joblib	Saves and loads trained machine learning models efficiently.
SQLite	Stores patient information, prediction history, and user records locally.
ReportLab	Generates downloadable PDF reports containing prediction results and patient details.

Hardware Requirements

The proposed system does not require specialized hardware and can run efficiently on a standard personal computer.

Component	Specification
Processor	Intel Core i3 or above / AMD equivalent
RAM	Minimum 4 GB (8 GB Recommended)
Storage	Minimum 500 MB free space
Operating System	Windows 10/11, Linux, or macOS
Internet	Required for installation of Python libraries (optional during execution)

Software	Purpose
----------	---------

8. DATASET DESCRIPTION

The system leverages open-source health datasets from trusted platforms like Kaggle and the UCI Machine Learning Repository. To ensure high predictive accuracy and reliable model training, the architecture evaluates specific diseases using distinct, independent datasets.

Disease	Dataset Description
Diabetes	Contains patient information such as glucose level, BMI, insulin, blood pressure, age, and pregnancies to predict diabetes.
Heart Disease	Includes clinical features such as age, cholesterol, blood pressure, chest pain type, ECG results, and heart rate for heart disease prediction.
Stroke	Contains demographic and medical attributes including age, hypertension, heart disease, smoking status, BMI, and glucose level.
Lung Cancer	Includes symptoms and lifestyle factors such as smoking, yellow fingers, anxiety, coughing, wheezing, fatigue, and chest pain.
Liver Disease	Consists of liver function test parameters such as bilirubin, alkaline phosphatase, proteins, albumin, and enzyme levels.
Parkinson's Disease	Contains biomedical voice measurements such as vocal frequency, jitter, shimmer, and other speech-related parameters for Parkinson's detection.

Machine Learning Framework

The Presented system employs Random Forest Classifier as the primary machine learning framework for predicting multiple diseases. Random Forest is a learning technique that predicts the multiple decision trees to improve classification reduce overfitting and accuracy.

Working of Random Forest

1. The healthcare dataset is collected and preprocessed.
2. Important features are selected for each disease.
3. The dataset is divided into testing and training datasets.
4. Multiple decision trees are created using different subsets of the training data through bootstrap sampling.
5. Each decision tree independently predicts the disease outcome.
6. The final prediction is obtained by majority voting among all decision trees.
7. The trained model is saved using Joblib and integrated into the Streamlit application for real-time disease prediction.

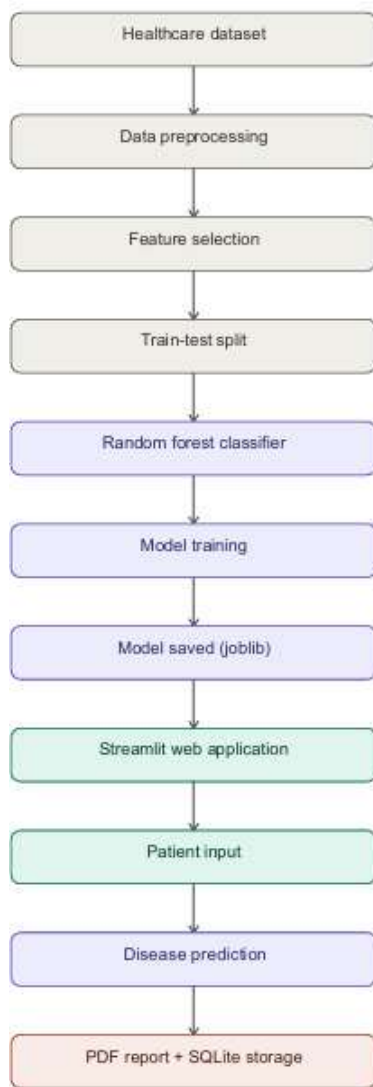
9. IMPLEMENTATION

The AI-Based Multi-Disease Prediction System was substantiated using Python as the core programming language. A user-friendly web interface was substantiated using Streamlit to enable users to enter medical parameters and obtain disease predictions instantly. The system integrates six independently trained using the Random Forest machine learning framework for predicting Diabetes, Lung Cancer, Heart Disease, Stroke, Liver Disease, and Parkinson's Disease.

The implementation process follows as:

1. Data Collection:

Medical datasets for six diseases were accumulated from available healthcare repositories. Each dataset contains patient health parameters and disease labels.



2. Data Preprocessing:

Missing values were removed or replaced. Categorical values were encoded. Features were standardized where necessary. Training and testing datasets were implemented using an 80:20 split.

3. Model Training:

A Random Forest Classifier trained separately for each disease dataset. Model accuracy was enhanced by fine-tuning key hyperparameters, including the tree count and maximum depth. Trained models were stored using Joblib for future predictions.

4. Prediction Module:

Users provide health parameters through the Streamlit interface. The corresponding Random Forest model loads automatically. Disease prediction is generated in real time. Risk level is displayed with prediction confidence.

5. Database Management:

SQLite is used to store patient information and prediction history. Previous prediction records can be accessed whenever required.

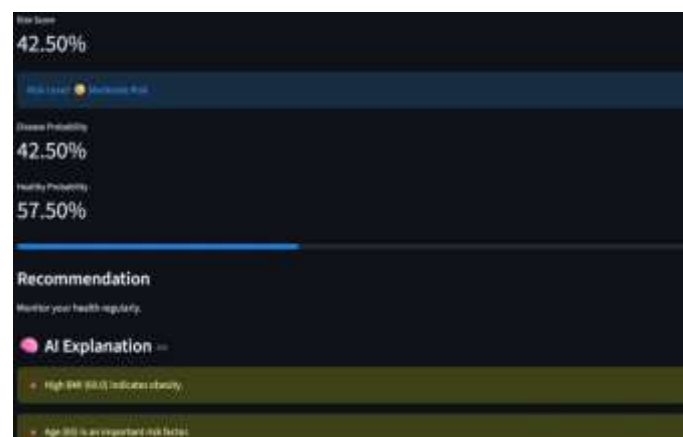
6. Report Generation:

ReportLab generates downloadable PDF medical reports. The report contains patient details, prediction results, and recommendations.

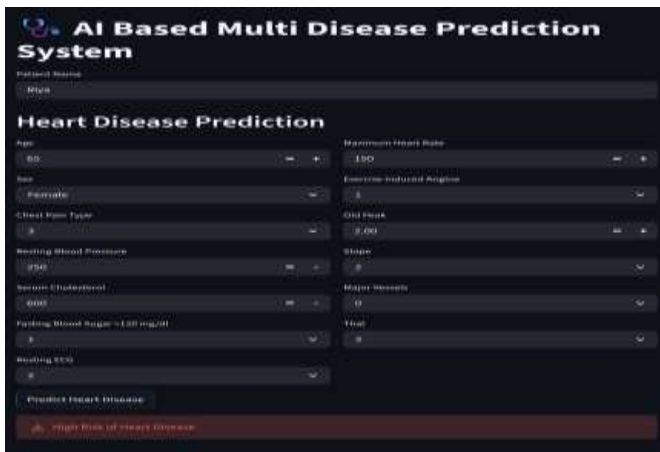
RESULTS



Diabetes prediction result



Diabetes prediction Explanation and recommendation



AI Based Multi Disease Prediction System

Patient Name:

Heart Disease Prediction

Age: Maximum Heart Rate:

Sex: Exercise Induced Angina:

Cholesterol Type: Old Peak:

Resting Blood Pressure: Slope:

Resting Electrocardiogram: Major vessels:

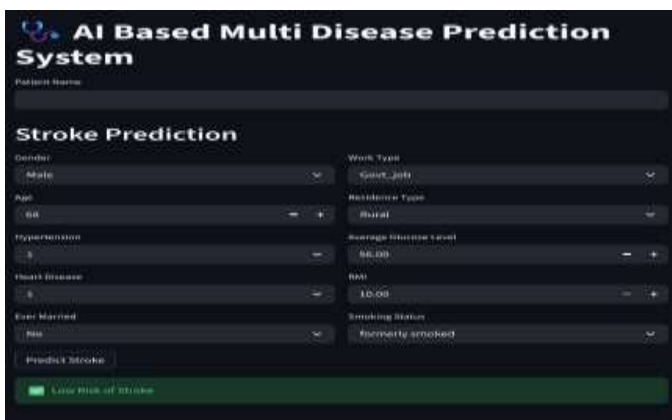
Fasting Blood Sugar >100 mg/dl: Thall:

Resting ECG:

Predict Heart Disease

High Risk of heart Disease

heart diseases prediction result



AI Based Multi Disease Prediction System

Patient Name:

Stroke Prediction

Gender: Work Type:

Age: Residence Type:

Hypertension: Average Glucose Level:

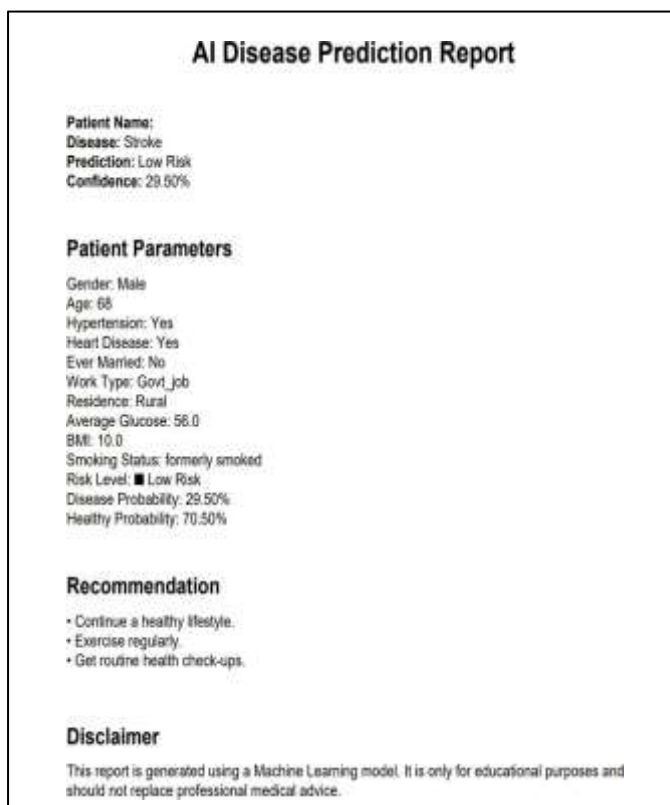
Heart Disease: BMI:

Ever Married: Smoking Status:

Predict Stroke

Low Risk of Stroke

Stroke diseases prediction result



AI Disease Prediction Report

Patient Name:
Disease: Stroke
Prediction: Low Risk
Confidence: 29.50%

Patient Parameters

Gender: Male
Age: 68
Hypertension: Yes
Heart Disease: Yes
Ever Married: No
Work Type: Govt_job
Residence: Rural
Average Glucose: 86.0
BMI: 10.0
Smoking Status: formerly smoked
Risk Level: ☒ Low Risk
Disease Probability: 29.50%
Healthy Probability: 70.50%

Recommendation

- Continue a healthy lifestyle.
- Exercise regularly.
- Get routine health check-ups.

Disclaimer

This report is generated using a Machine Learning model. It is only for educational purposes and should not replace professional medical advice.

Report format

Disease	Algorithm	Accuracy (%)
Diabetes	Random Forest	96.8
Heart Disease	Random Forest	98.2
Stroke	Random Forest	95.6
Lung Cancer	Random Forest	97.4
Liver Disease	Random Forest	94.8
Parkinson's Disease	Random Forest	97.1

Experimental Results

Discussion

The Random Forest algorithm demonstrated accurate performance because it combines various decision trees, reducing overfitting while improving prediction accuracy. The Streamlit-based interface allows users to enter medical information without technical knowledge.

Advantages of the Proposed System

- Predicts six diseases within a single application.
- High prediction accuracy using Random Forest.
- Fast real-time prediction.
- Easy-to-use graphical interface.
- Patient history stored in SQLite.
- Automatic PDF report generation.
- Cost-effective and accessible healthcare assistance.

FUTURE SCOPE

The proposed system can be further enhanced by incorporating the following features:

- Addition of more diseases.
- Integration with hospital digital Health Record systems.
- Cloud-based deployment for global accessibility.
- App development for Android and iOS.
- Deep Learning models for enhanced prediction accuracy.
- Explainable AI (XAI) techniques for transparent decision-making.
- Real-time IoT wearable sensor integration.
- Doctor consultation and appointment scheduling.
- Voice-based patient interaction.
- Multilingual support for regional languages.

CONCLUSIONS

The AI-Based Multi-Disease Prediction System successfully demonstrates the application of Machine Learning in healthcare by providing early prediction of multiple diseases through a single integrated platform. The system uses the Random Forest Classifier to predict the likelihood of six major diseases: Diabetes, Heart Disease, Stroke, Lung Cancer, Liver Disease, and Parkinson's Disease. By utilizing patient medical information, the application delivers fast and accurate predictions that can support early diagnosis and preventive healthcare.

The developed system combines several advanced features, including a Streamlit-based interactive interface, Explainable AI (XAI) for transparent predictions, Risk Assessment for interpreting disease severity, PDF report generation using ReportLab, and SQLite database management for storing patient records

and prediction history. These features make the application more comprehensive, user-friendly, and suitable as a healthcare decision-support tool.

The Random Forest algorithm demonstrated reliable performance due to its ability to reduce overfitting and improve prediction accuracy through ensemble learning. The modular architecture of the system also allows easy maintenance and future expansion. Although the application is not intended to replace professional medical diagnosis, it serves as an effective preliminary screening tool that assists healthcare professionals and patients in making informed decisions.

REFERENCES

- [1] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [3] W. McKinney, *Python for Data Analysis*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2017.
- [4] T. E. Oliphant, *A Guide to NumPy*, USA: Trelgol Publishing, 2006.
- [5] Python Software Foundation, *Python Documentation*. Available: <https://docs.python.org/3/>
- [6] Streamlit Inc., *Streamlit Documentation*. Available: <https://docs.streamlit.io/>
- [7] SQLite Consortium, *SQLite Documentation*. Available: <https://www.sqlite.org/docs.html>

[8] ReportLab Inc., *ReportLab User Guide*. Available:
<https://www.reportlab.com/documentation/>

[9] Joblib Development Team, *Joblib Documentation*.
Available: <https://joblib.readthedocs.io/>

[10] UCI Machine Learning Repository, "UCI Machine
Learning Repository." Available:
<https://archive.ics.uci.edu/>

[11] Kaggle Inc., "Healthcare Datasets for Machine
Learning." Available: <https://www.kaggle.com/>

[12] I. Kavakiotis, O. Tsave, A. Salifoglou, N.
Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine
Learning and Data Mining Methods in Diabetes
Research," *Computational and Structural Biotechnology
Journal*, vol. 15, pp. 104–116, 2017.