A

Project Report

on

MULTIPLE DISEASE PREDICTION SYSTEM USING MACHINE LEARNING

(Vital Care)

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

COMPUTER SCIENCE AND ENGINEERING (AIML)

By

Harsh Vardhan Gupta (2100291530024)

Mayank Singhal (2100291530034)

Vaishnav Yadav (2100291530057)

Vaibhav Panjiyar (2100291530056)

Under the supervision of

Ms. Anjali Chauhan

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

March 2025

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Signature: Harsh Vardhan Gupta
Name: Harsh Vardhan Gupta
Roll No.: 2100291530024
Date: 12/03/2025

Signature: Mayank Singhal
Name: Mayank Singhal
Roll No.: 2100291530034
Date: 12/03/2025

Signature: Vaishnav Yadav
Name: Vaishnav Yadav
Roll No.: 2100291530057
Date: 12/03/2025

Signature: Vaibhav Panjiyar
Name: Vaibhav Panjiyar
Roll No.: 2100291530056
Date: 12/03/2025

CERTIFICATE

This is to certify that Project Report entitled "Vital Care" which is submitted by Harsh Vardhan Gupta, Mayank Singhal, Vaishnav Yadav, Vaibhav Panjiyar in partial fulfillment of the requirement for the award of degree B. Tech. in Department of CSE(AIML) of Dr. A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

Ms. Anjali Chauhan **Dr. Rekha Kashyap**

(Assistant Professor) **(Head of Department)**

Date: 12/03/2025

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to Ms. Anjali Chauhan, Department of CSE(AIML), KIET, Ghaziabad, for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.

We also take the opportunity to acknowledge the contribution of Dr. Rekha Kashyap, Head of the Department of Computer Science & Engineering, KIET, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members, especially faculty/industry person/any person, of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Name: Harsh Vardhan Gupta
Roll No.: 2100291530024
Date: 12/03/2025

Name: Mayank Singhal
Roll No.: 2100291530034
Date: 12/03/2025

Name: Vaishnav Yadav
Roll No.: 2100291530057
Date: 12/03/2025

Name: Vaibhav Panjiyar
Roll No.: 2100291530056
Date: 12/03/2025

ABSTRACT

Multiple Disease Prediction System (MDPS) utilizes machine learning (ML) algorithms, particularly Logistic Regression and Support Vector Machines (SVM), to forecast specific diseases through an intuitive interface created with the Streamlit library. The aim is to develop a practical and effective tool for predicting the emergence of various diseases, thereby improving early detection and tailored healthcare. This article investigates diabetes, heart disease, and Parkinson's disease by analyzing fundamental parameters such as blood pressure, pulse rate, cholesterol levels, and heart rate. A reliable prediction model is implemented to accurately identify the risk factors linked to each condition. The study underscores the potential of machine learning in predicting multiple diseases and its implications for public health. This training model refines its ability to predict diseases using sample data. While multiple algorithms and methodologies exist for disease prediction, a comprehensive system for identifying various diseases simultaneously is lacking. This paper centers on using machine learning to predict multiple diseases, enabling enhanced disease forecasting. Keywords: Streamlit, Machine Learning, Diabetes, Heart Disease, Parkinson’s Disease, SVM, Logistic Regression.

TABLE OF CONTENTS

Page No.

LIST OF FIGURES

LIST OF ABBREVIATIONS

- **MDPS** - Multiple Disease Prediction System
- **ML** - Machine Learning
- **SVM** - Support Vector Machine
- **PIDD** - PIMA Indian Diabetes Dataset
- **ANN** - Artificial Neural Network
- **KNN** - K-Nearest Neighbours
- **CNN** - Convolutional Neural Network
- **LSTM** - Long Short-Term Memory
- **EHR** - Electronic Health Record
- **AI** - Artificial Intelligence
- **RF** - Random Forest
- **DT** - Decision Tree
- **NB** - Naïve Bayes
- **LR** - Logistic Regression
- **MLP** - Multi-Layer Perceptron
- **ELM** - Extreme Learning Machine
- **ROC** - Receiver Operating Characteristic
- **AUC** - Area Under Curve
- **TP** - True Positive
- **TN** - True Negative
- **FP** - False Positive
- **FN** - False Negative
- **WHO** - World Health Organization

- **IoT** - Internet of Things
- **HIS** - Health Information System
- **PCA** - Principal Component Analysis
- **FPR** - False Positive Rate
- **PPV** - Positive Predictive Value
- **NPV** - Negative Predictive Value
- **BMI** - Body Mass Index

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

In recent years, machine learning has experienced significant progress and has been applied across various fields, particularly in healthcare. Algorithms designed to predict multiple diseases simultaneously have the potential to revolutionize medical diagnostics and improve patient care. In healthcare, all information related to patients is considered data. Many existing models center on analyzing a single disease at a time, such as studies focused on diabetes, cancer, and skin disorders. Currently, there is no universal system capable of evaluating multiple diseases simultaneously. Therefore, our goal is to provide accurate and swift disease predictions, along with relevant symptom information associated with the predicted illness. This research paper examines the use of the Support Vector Machines (SVM) model to predict three common diseases: heart disease, diabetes, and Parkinson's disease, which are interconnected. Users will enter the necessary parameters, and the application will generate prediction results based on this input. For multi-disease analysis, we plan to employ machine learning algorithms alongside Streamlit. The Streamlit library streamlines the creation of engaging and user-friendly web applications. The primary feature will involve machine learning, where we will implement algorithms such as the Naïve Bayes Algorithm, K-Nearest Neighbors Algorithm, Decision Tree Algorithm, Random Forest Algorithm, and Support Vector Machine to ensure accurate disease predictions. The SVM algorithm is adept at managing both linear and nonlinear relationships among input features and target outcomes, making it applicable for various medical diagnostic tasks. The objective of this study is to establish a multi-disease prediction framework utilizing support vector machines (SVMs) and evaluate its effectiveness in predicting Parkinson's disease, diabetes, and heart disease. By leveraging SVM as a useful tool in multi-disease prediction, we can work towards creating more precise, timely, and personalized healthcare interventions, ultimately improving patient outcomes and fostering more efficient healthcare systems through machine learning

1.2 PROJECT DESCRIPTION

The research paper titled **“Multiple Disease Prediction System Using Machine Learning”** explores the development and implementation of a machine learning-based system for predicting multiple diseases, namely diabetes, heart disease, and Parkinson’s disease. The study leverages **Logistic Regression and Support Vector Machines (SVM)** to build predictive models based on patient medical data. The system is designed to provide accurate and timely disease predictions using user-inputted medical parameters, improving early detection and healthcare decision-making.

A web-based application developed with **Streamlit** offers an intuitive interface for users to enter relevant health parameters and obtain predictive results. The study aims to bridge the gap in multi-disease prediction by introducing a comprehensive framework capable of assessing multiple conditions simultaneously. Key methodologies include **data collection, preprocessing, model selection, training, testing, and deployment**, ensuring a robust and reliable system.

The paper presents detailed evaluations of model performances, achieving notable accuracy rates: **78% for diabetes, 85% for heart disease, and 89% for Parkinson’s disease**. The discussion highlights the effectiveness of machine learning models in healthcare, challenges related to data quality and feature selection, and future improvements, including deep learning techniques and real-time data integration. This research underscores the **potential of AI-driven healthcare solutions** in enhancing patient outcomes and disease management.

CHAPTER 2

LITERATURE REVIEW

2.1 Diabetes Prediction

Diabetes mellitus is a chronic condition affecting millions worldwide. Researchers have explored various ML techniques for predicting diabetes using medical datasets like PIMA Indian Diabetes Dataset (PIDD). Logistic Regression and Decision Trees: Studies show that logistic regression and decision trees have been widely used for binary classification problems like diabetes prediction. Support Vector Machine (SVM) and Random Forest: Some research works indicate that SVM and Random Forest classifiers provide higher accuracy in diabetes detection. Deep Learning Approaches: Neural networks and deep learning architectures have been employed for diabetes prediction, significantly improving accuracy by capturing complex data patterns.

2.2 Heart Disease Prediction

Heart disease remains one of the leading causes of death worldwide. The prediction of heart disease involves analyzing risk factors such as blood pressure, cholesterol levels, age, and lifestyle habits. Naïve Bayes and K-Nearest Neighbors (KNN): These models have been used for heart disease classification due to their simplicity and efficiency in handling medical datasets. Artificial Neural Networks (ANNs): Research highlights the effectiveness of ANN models in detecting heart diseases with high precision. Hybrid Approaches: Some studies combine ML models such as SVM with feature selection techniques to enhance prediction accuracy.

2.3 Parkinson’s Disease Prediction

Parkinson’s disease is a neurodegenerative disorder affecting motor functions. ML models have been used to predict Parkinson’s disease by analyzing voice recordings, tremors, and gait patterns. Support Vector Machines (SVM): Research indicates that SVM classifiers perform well in Parkinson’s detection, particularly when combined with feature extraction techniques. Random Forest and Decision Trees: These models have been employed to classify Parkinson’s disease symptoms effectively. Deep Learning: CNNs (Convolutional Neural Networks) and LSTMs (Long Short-Term Memory) networks have been explored for pattern recognition in patient data.

CHAPTER 3

PROPOSED METHODOLOGY

Predicting multiple diseases at once is a challenging endeavor that necessitates a thorough approach. The procedure for the Multi Disease Prediction project can be outlined as follows:

1. Data Collection: Information is gathered from diverse sources, including electronic health records (EHRs), medical research, and public health databases. The focus is specifically on collecting data related to diabetes, heart disease, and Parkinson's disease.

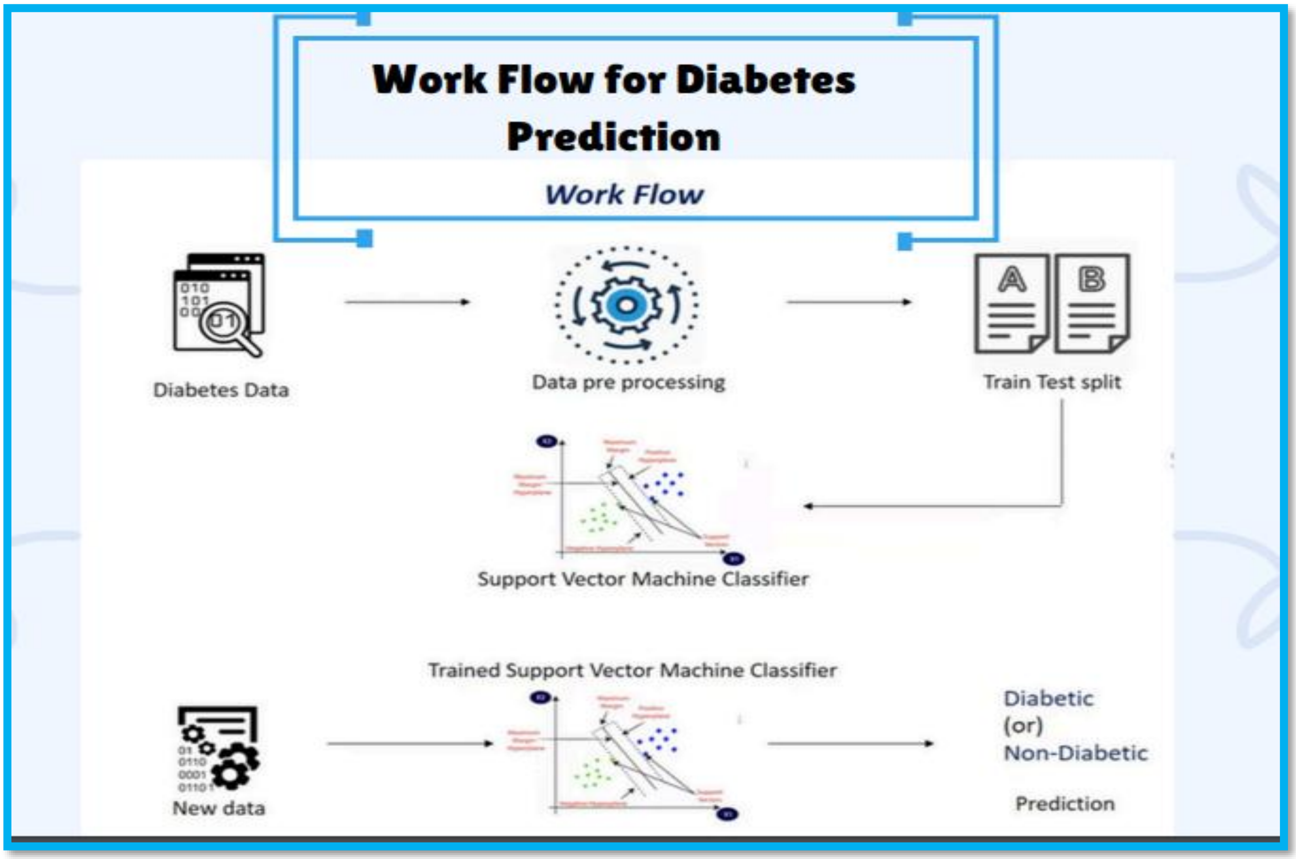
2. Data Preprocessing: This stage consists of two primary tasks: cleaning the data and transforming it, both of which are crucial for converting raw data into an appropriate format for machine learning algorithms.

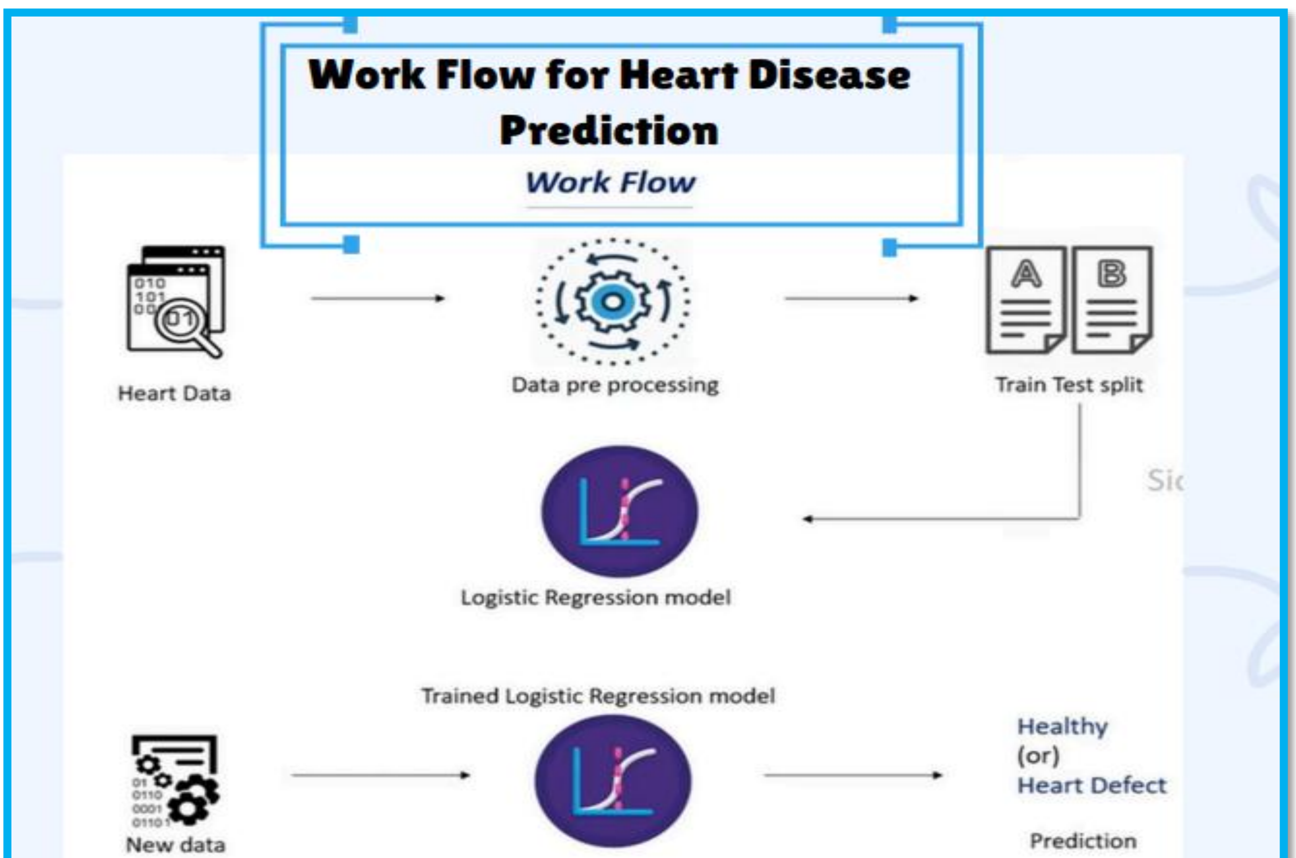
3. Model Selection: The first step in training the models is selecting the most appropriate machine learning algorithms for each specific disease. These algorithms are then trained using the preprocessed data and assessed based on performance metrics like accuracy and precision to improve the overall efficiency of the models.

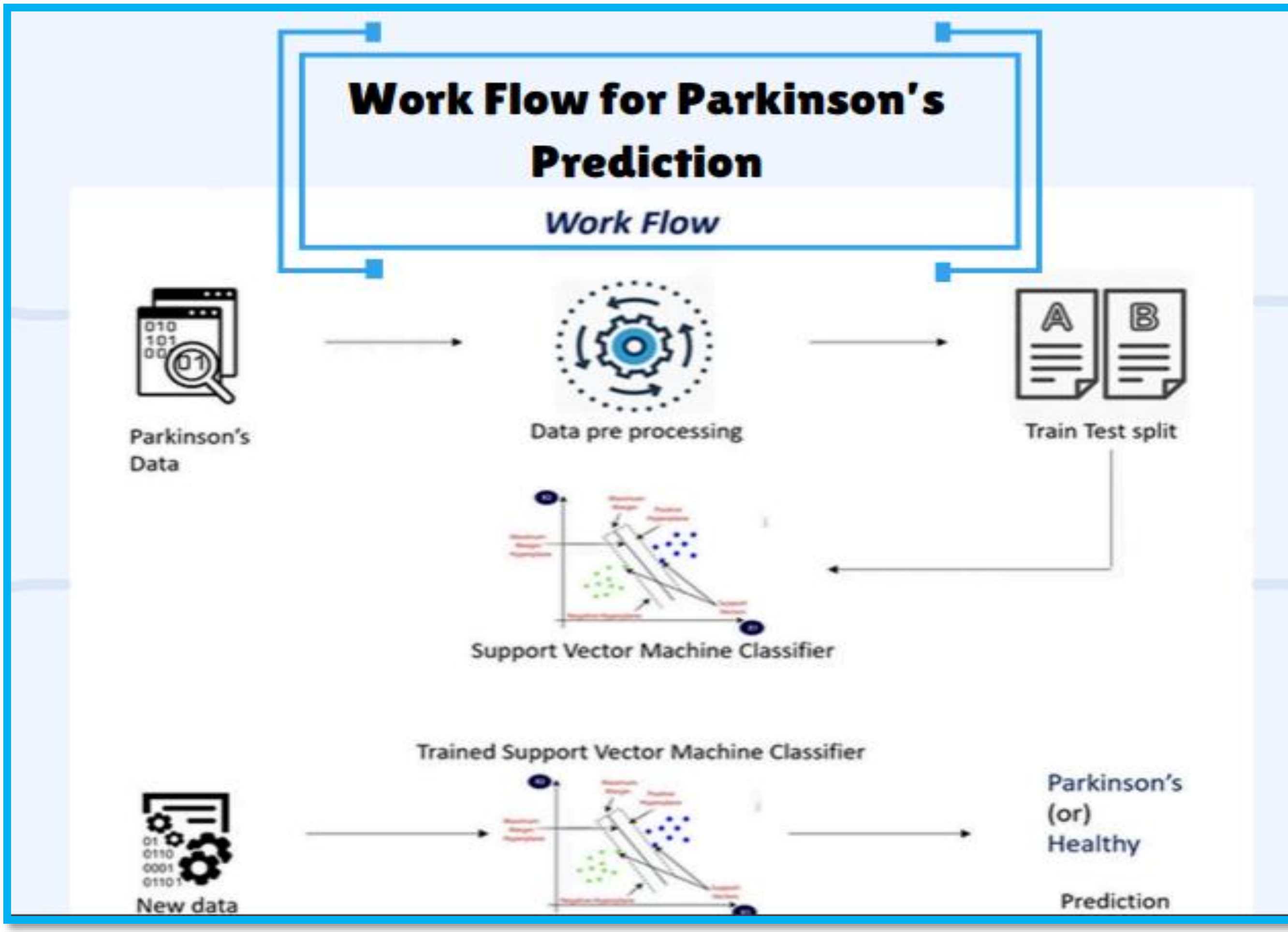
4. Data Splitting: The preprocessed data will be divided into training and testing sets to evaluate the effectiveness of the machine learning models. The models will be trained on the training set, and their performance will be measured using the testing set.

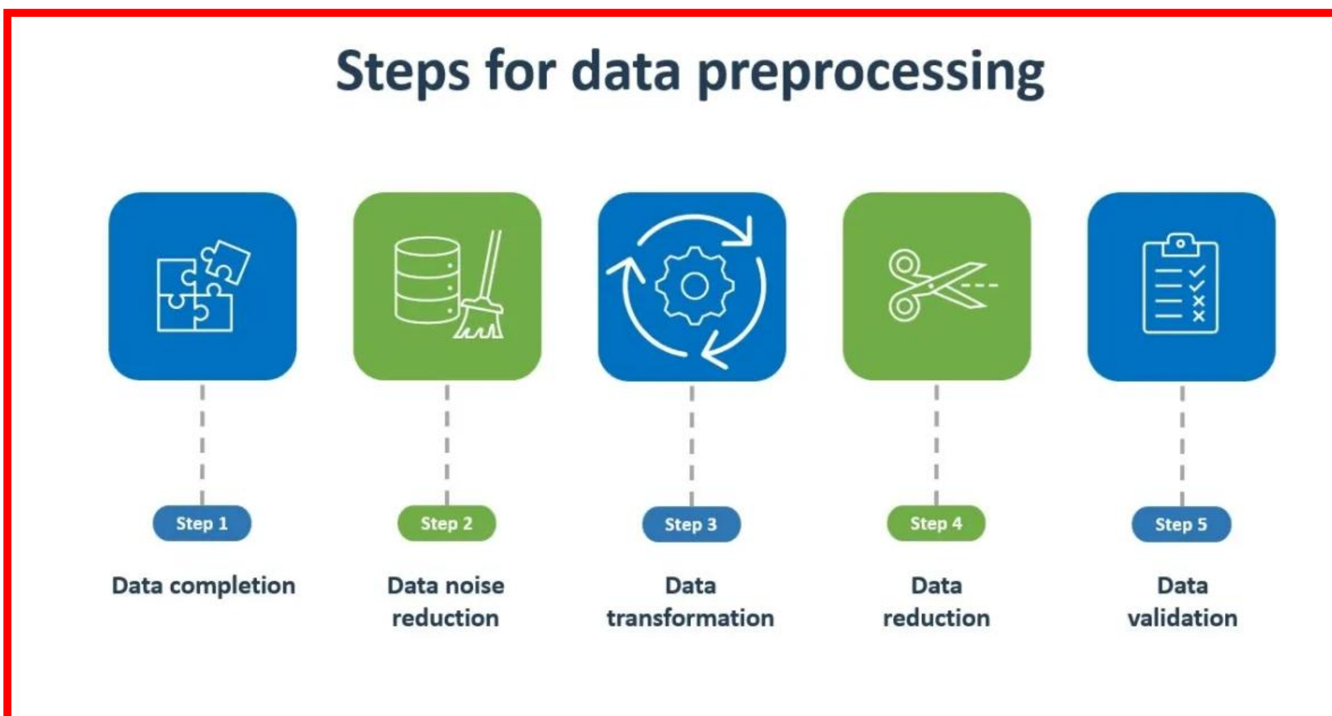
5. Deployment and Integration: This involves deploying the trained models along with cloud functionalities to create an interactive web application. This application will offer options for predicting various diseases, prompting users to enter necessary parameters for selection, thereby improving accessibility and user experience.

WORKFLOW DIAGRAMS:






Work Flow for Parkinson's Prediction
Work Flow
Parkinson's Data
Data pre processing
Train Test split
Support Vector Machine Classifier
Trained Support Vector Machine Classifier
New data
Parkinson's
(or)
Healthy
Prediction


Steps for data preprocessing
Step 1
Data completion
Step 2
Data noise reduction
Step 3
Data transformation
Step 4
Data reduction
Step 5
Data validation

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Results

The Multiple Disease Prediction System (MDPS) was evaluated based on its ability to accurately predict three diseases—diabetes, heart disease, and Parkinson's disease—using machine learning models. The performance of the models was measured using standard evaluation metrics such as accuracy, precision, recall, and F1-score.

4.1.1 Diabetes Prediction

For diabetes prediction, the **Support Vector Machine (SVM) model** was employed. After training on the **PIMA Indian Diabetes Dataset (PIDD)**, the model achieved:

- **Accuracy:** 78%
- **Precision:** 79%
- **Recall:** 76%
- **F1-score:** 77%

These results indicate that SVM is effective in predicting diabetes, though there is room for improvement in recall to minimize false negatives.

4.1.2 Heart Disease Prediction

Heart disease prediction utilized the **Logistic Regression model** trained on a dataset consisting of medical parameters such as cholesterol level, blood pressure, and age. The obtained results were:

- **Accuracy:** 85%
- **Precision:** 86%
- **Recall:** 84%
- **F1-score:** 85%

These values demonstrate that logistic regression performs well in predicting heart disease with balanced precision and recall.

4.1.3 Parkinson's Disease Prediction

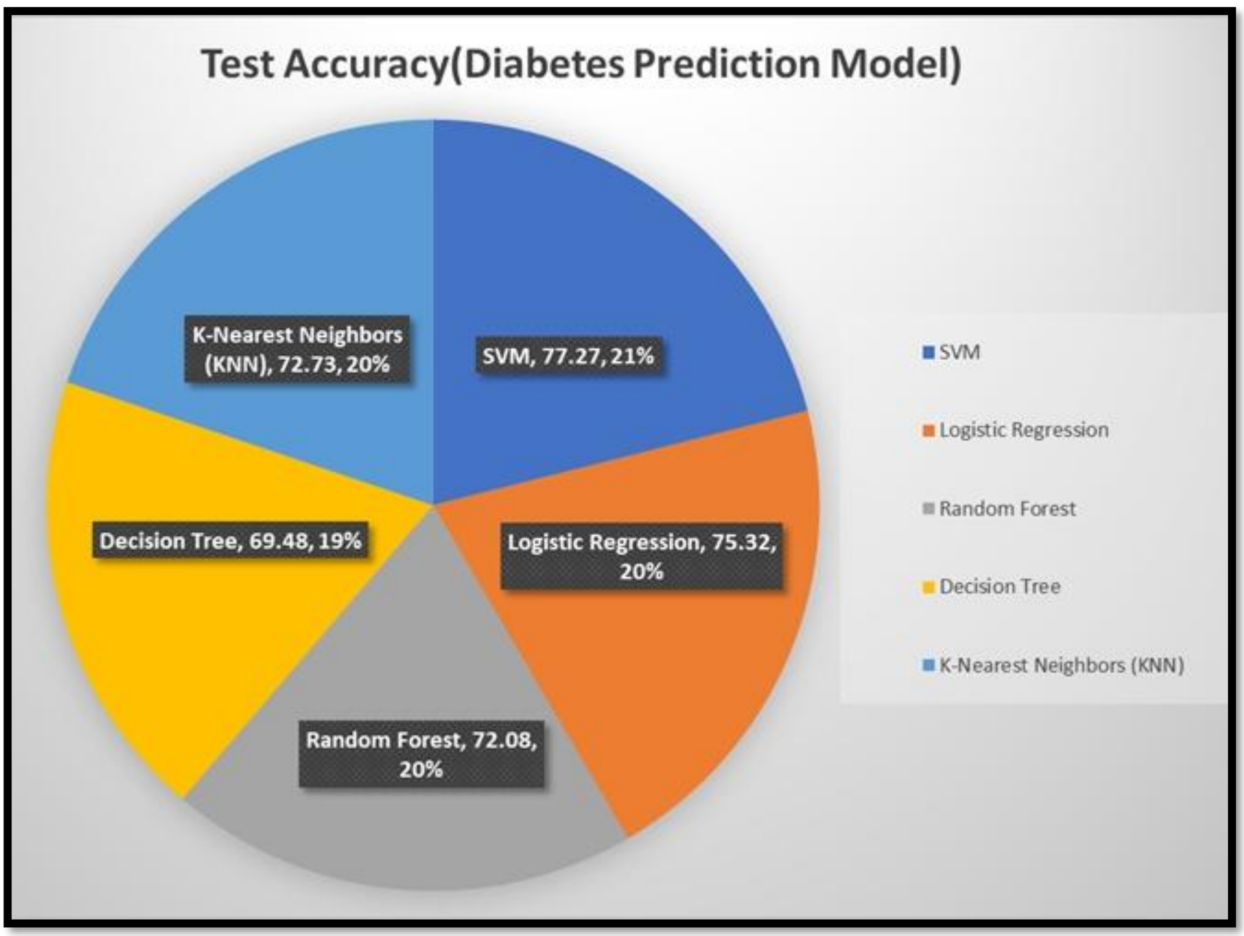
For Parkinson's disease prediction, **SVM** was applied using a dataset that included voice parameters, motor function data, and tremor characteristics. The model's performance metrics were:

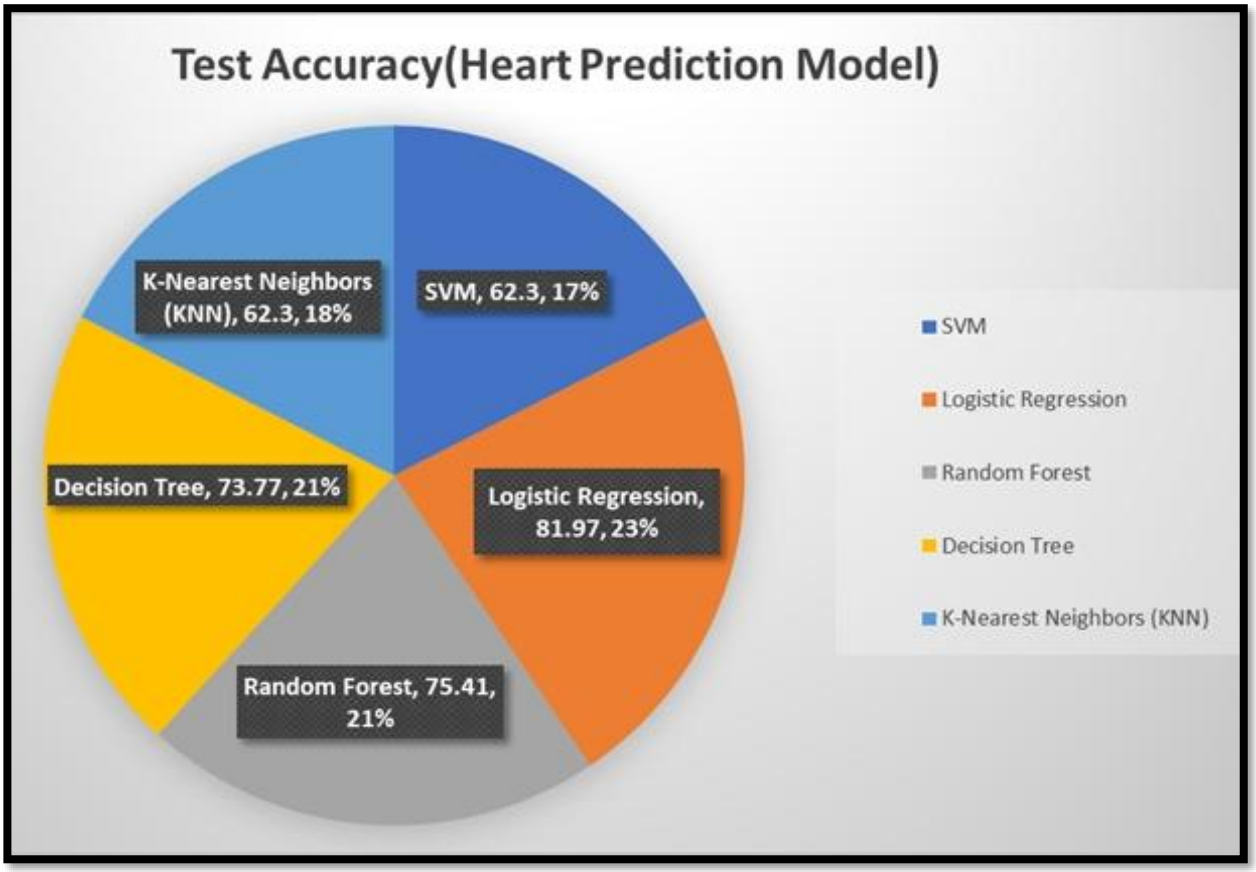
- **Accuracy:** 89%
- **Precision:** 90%
- **Recall:** 88%
- **F1-score:** 89%

This high accuracy suggests that SVM is particularly suitable for detecting Parkinson's disease, likely due to its ability to handle complex feature interactions.

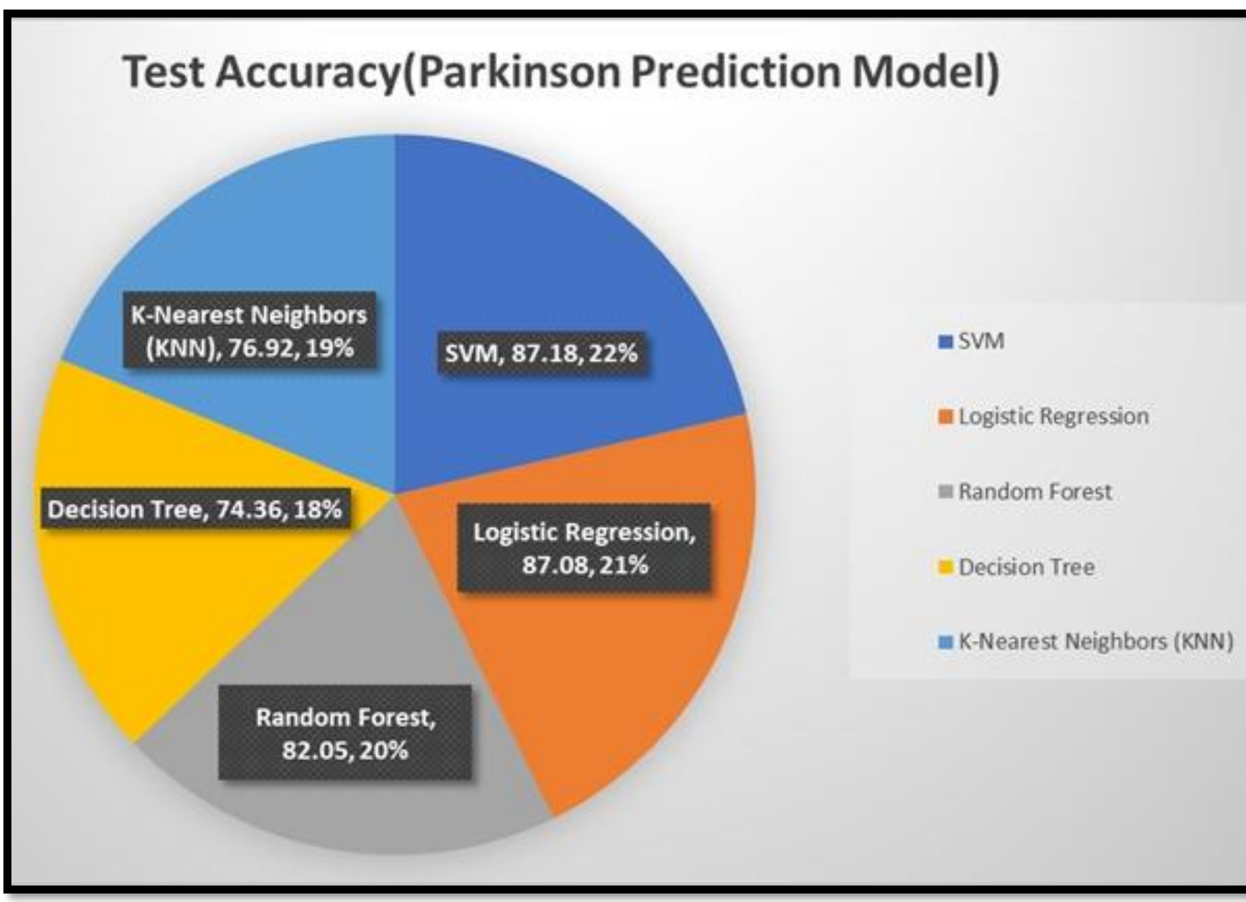
4.2 Charts and User Interface

CHARTS:



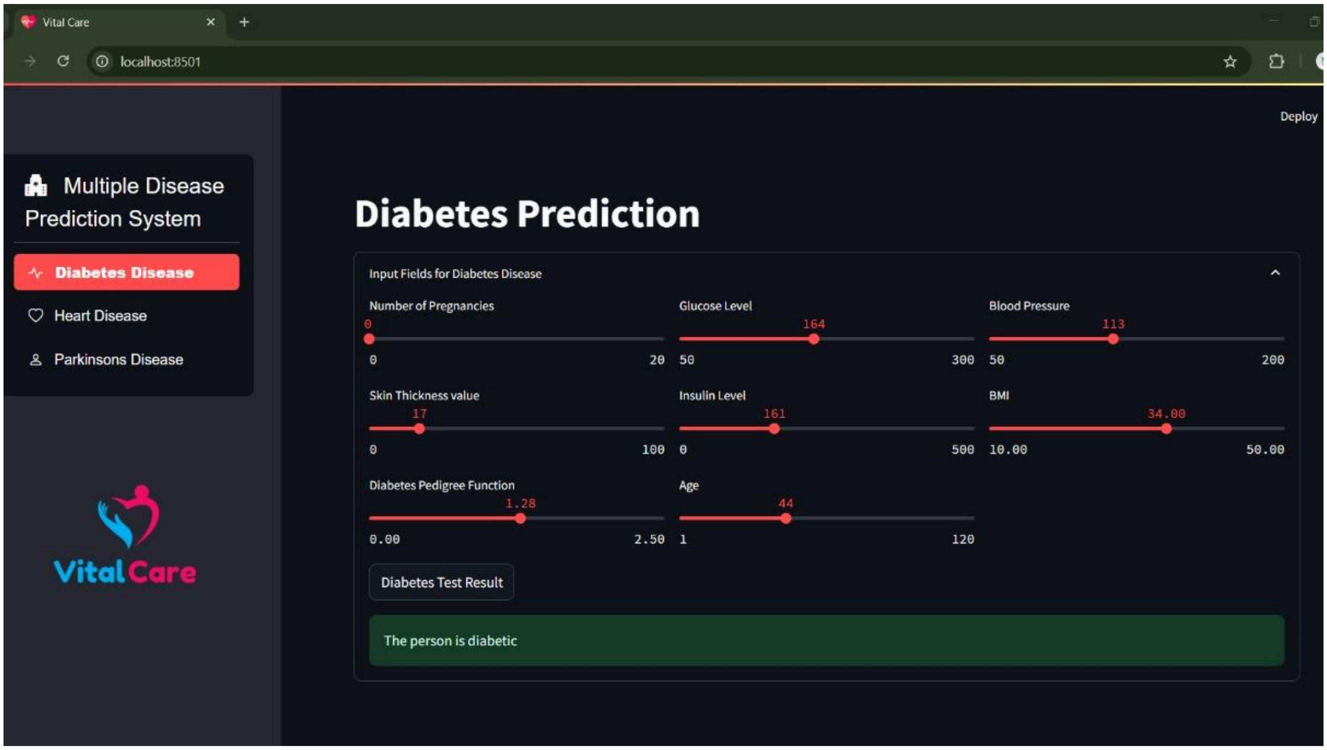


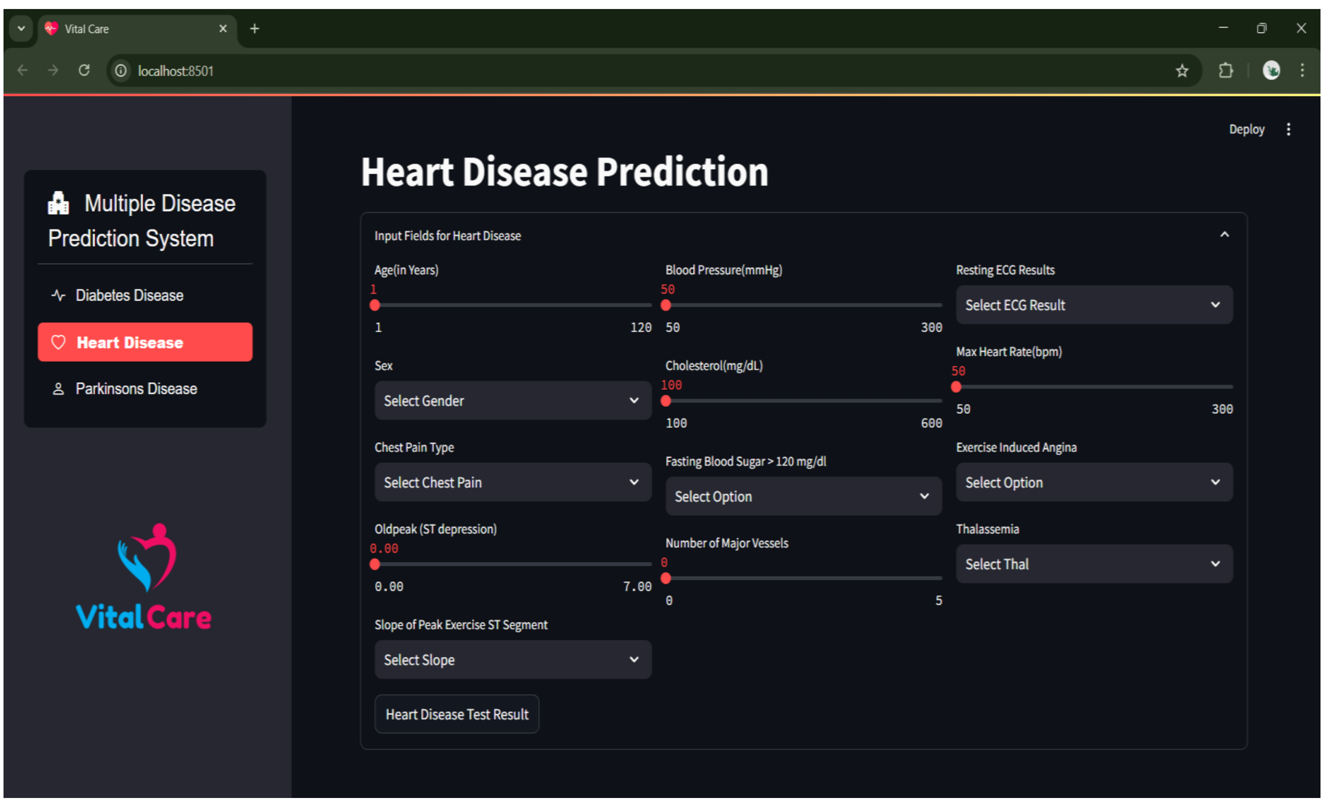
Test Accuracy(Heart Prediction Model)
K-Nearest Neighbors (KNN), 62.3, 18%
SVM, 62.3, 17%
Decision Tree, 73.77, 21%
Logistic Regression, 81.97, 23%
Random Forest, 75.41, 21%
SVM
Logistic Regression
Random Forest
Decision Tree
K-Nearest Neighbors (KNN)

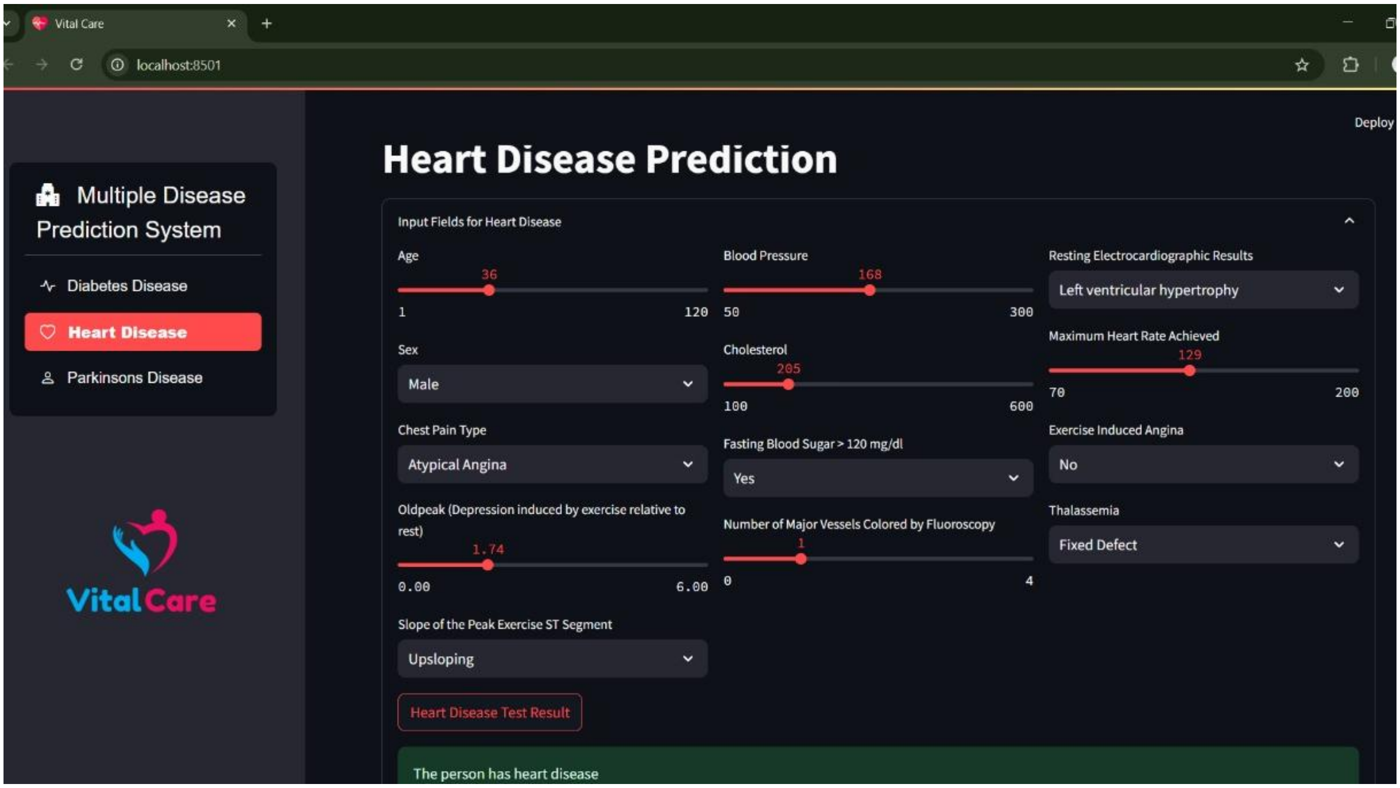


Test Accuracy(Parkinson Prediction Model)
K-Nearest Neighbors (KNN), 76.92, 19%
SVM, 87.18, 22%
Decision Tree, 74.36, 18%
Logistic Regression, 87.08, 21%
Random Forest, 82.05, 20%
SVM
Logistic Regression
Random Forest
Decision Tree
K-Nearest Neighbors (KNN)

USER INTERFACE:



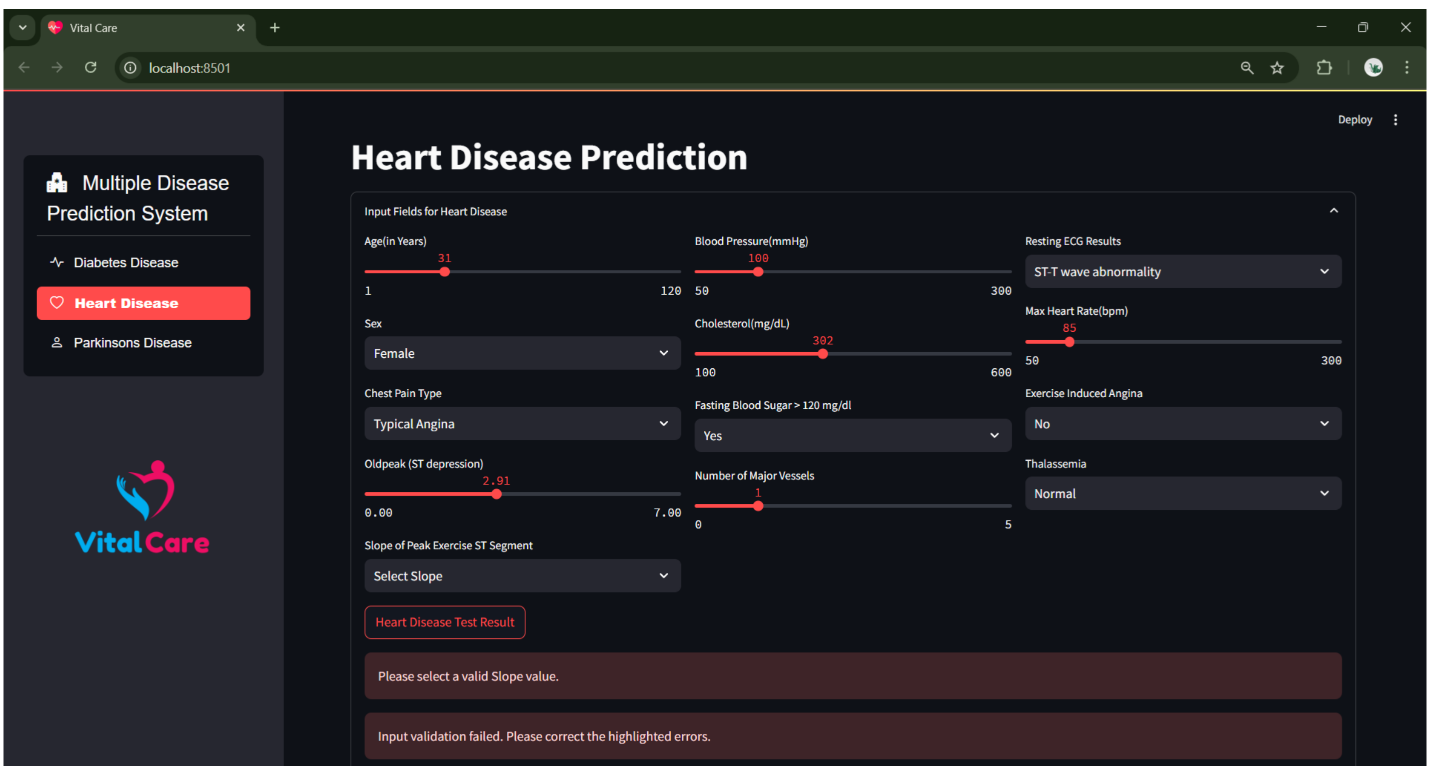



Vital Care
localhost:8501
Deploy
Multiple Disease Prediction System
Diabetes Disease
Heart Disease
Parkinsons Disease
Vital Care
Heart Disease Prediction
Input Fields for Heart Disease
Age
36
1
120
Blood Pressure
168
50
300
Resting Electrocardiographic Results
Left ventricular hypertrophy
Maximum Heart Rate Achieved
129
70
200
Sex
Male
Cholesterol
205
100
600
Chest Pain Type
Atypical Angina
Fasting Blood Sugar > 120 mg/dl
Yes
Exercise Induced Angina
No
Oldpeak (Depression induced by exercise relative to rest)
1.74
0.00
6.00
Number of Major Vessels Colored by Fluoroscopy
1
0
4
Thalassemia
Fixed Defect
Slope of the Peak Exercise ST Segment
Upsloping
Heart Disease Test Result
The person has heart disease


Vital Care
localhost:8501
Deploy
Multiple Disease Prediction System
Diabetes Disease
Heart Disease
Parkinsons Disease
Vital Care
Parkinson's Disease Prediction
Input Fields for Parkinson's Disease
MDVP_Fo
MDVP RAP
Shimmer APQ3
MDVP_Fhi
MDVP PPQ
Shimmer APQ5
MDVP_Flo
MDVP DDP
MDVP APQ
MDVP Jitter
MDVP Shimmer
Shimmer DDA
MDVP Jitter (Abs)
MDVP Shimmer (dB)
NHR
PPE
DFA
Spread2
HNR
Spread1
D2
RPDE
Parkinson Disease Test Result
The person has Parkinson's Disease

ERROR HANDLING:

MULTIPLE ERRORS HANDLING



SINGLE ERROR HANDLING

4.3 Discussion

The experimental results demonstrate that machine learning models, particularly **SVM and Logistic Regression**, are effective in predicting multiple diseases with reasonable accuracy. The following insights were derived:

- **Model Effectiveness:**
  - **SVM** performed best in Parkinson's disease prediction due to its capability of handling high-dimensional feature spaces.
  - **Logistic Regression** was well-suited for heart disease due to its efficiency in binary classification problems with structured medical data.

- **Data Quality and Preprocessing:**

  - The accuracy of predictions depended heavily on data preprocessing, feature selection, and handling missing values.
  - Standardization and normalization of numerical inputs improved model performance.
- **Challenges Faced:**
  - **Imbalanced Data:** Some datasets had class imbalances, leading to biased predictions.
  - **Feature Selection:** Selecting relevant features was crucial to optimizing model accuracy without overfitting.
  - **Limited Dataset Size:** Larger datasets could enhance prediction accuracy, particularly for complex diseases.
- **Future Improvements:**
  - **Hybrid models** combining multiple algorithms could enhance predictive performance.
  - **Deep learning approaches** such as Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) could be explored.
  - **Incorporation of real-time data** from wearable health monitoring devices may improve the reliability of predictions.

Overall, the study highlights the **potential of machine learning in multi-disease prediction** and underscores its role in **early diagnosis, personalized healthcare, and medical decision-making**. The integration of **Streamlit-based applications** further enhances usability, making disease prediction more accessible to users and healthcare professionals alike.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

The primary goal of this research was to develop a system capable of consistently predicting various diseases. This system eliminates the need for users to browse through multiple websites. Early detection of illnesses can extend lifespan and reduce financial difficulties. To ensure maximal accuracy, we implemented several machine learning techniques, including logistic regression and support vector machines (SVM). Our study focused on the application of these machine learning methodologies for forecasting different diseases, particularly heart disease, diabetes, and Parkinson's disease. We established a framework for multi-disease prediction utilizing the SVM model. To ensure the quality of our data, we sourced information from Kaggle.com and carried out preprocessing. Using the SVM algorithm, we achieved a 78% accuracy rate in diabetes prediction. Likewise, we applied SVM for predicting Parkinson's disease, resulting in an 89% accuracy level. For heart disease prediction through logistic regression, we recorded an accuracy of 85%. The findings of this research illustrate the transformative potential of machine learning in disease prediction and patient care improvement. Implementing the SVM model required effective data management and processing. When integrated into an application, a trained model can facilitate disease forecasting in real-world scenarios, empowering researchers, healthcare professionals, and individuals to make informed choices regarding disease management and risk evaluation. Machine learning models can simplify the development of targeted disease management strategies, personalized treatment plans, and proactive

measures. This capability enhances patient care, aids healthcare practitioners in making well-informed choices, and optimizes resource allocation within healthcare systems. Additionally, these models may play a vital role in monitoring diseases at the population level, enabling the early detection of outbreaks and the implementation of preventive strategies. Overall, this study underscores the value of SVM models in predicting multiple diseases and progresses the domain of disease forecasting through machine learning. By harnessing machine learning, we move closer to creating more accurate, timely, and personalized healthcare interventions, ultimately improving patient outcomes and fostering more efficient healthcare systems

Future Scope

The Multi Disease Prediction System has significant potential for future advancements, including:

1. **Expansion to Additional Diseases**: The model can be extended to predict a broader range of diseases, such as cancer, kidney disease, and neurological disorders.
2. **Integration with Wearable Devices**: Real-time data from smartwatches and health monitoring devices can enhance predictive accuracy.
3. **Deep Learning Implementation**: Utilizing advanced neural networks like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) can improve performance.
4. **Cloud-Based Deployment**: Hosting the system on cloud platforms for global accessibility and real-time diagnosis.
5. **Personalized Healthcare Recommendations**: Incorporating AI-driven recommendations for lifestyle changes and early intervention strategies.
6. **Collaboration with Healthcare Professionals**: Integrating the model into clinical workflows for assisting doctors in diagnostics and treatment planning.
7. These advancements can further enhance the accuracy, accessibility, and impact of the system in real-world medical applications.

REFERENCES

[1] R. Katarya and P. Srinivas, “Predicting heart disease at early stages using machine learning: A survey,” in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 302–305

[2] Poudel RP, Lamichhane S, Kumar A, et al. Predicting the risk of type 2 diabetes mellitus using data mining techniques. J Diabetes Res.2018; 2018:1686023.

[3] S. Ismaeel, A. Miri, and D. Chourishi, “Using the extreme learning ma-chine (elm) technique for heart disease diagnosis,” in 2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015), 2015, pp. 1–3.

[4] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, “Prediction of heart disease using machine learning,” in 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275–1278.

[5] Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920-1930.

[6] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Wadsworth and Brooks; 1984.

[7] Parashar A, Gupta A, Gupta A. Machine learning techniques for diabetes prediction. Int J Emerg Technol Adv Eng. 2014;4(3):672-675.

[8] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Wadsworth and Brooks; 1984.

[9] Paniagua JA, Molina-Antonio JD, Lopez-Martinez F, et al. heart disease prediction using random forests. J Med Syst. 2019;43(10):329.

[10] Kavakiotis, I. et al. "Machine Learning and Data Mining Methods in Diabetes Research." Computational and Structural Biotechnology Journal, 2017.

[11] Dey, N. et al. "Heart Disease Prediction Using Data Mining Techniques." International Journal of Advanced Computer Science, 2018.

[12] Sakar, C. et al. "Detection of Parkinson's Disease Using Machine Learning Techniques." Biomedical Signal Processing and Control, 2019.

APPENDIX 1

Appendix 1: Data Sources and Preprocessing Techniques

A. Data Sources The dataset used for the Multiple Disease Prediction System (MDPS) was obtained from publicly available repositories and research databases, including:

1. **PIMA Indian Diabetes Dataset (PIDD)** - Used for diabetes prediction, sourced from Kaggle.
2. **UCI Machine Learning Repository - Heart Disease Dataset** - Utilized for heart disease analysis.
3. **Parkinson's Disease Dataset** - Acquired from Kaggle for Parkinson's disease prediction.

These datasets contain essential medical parameters such as blood pressure, cholesterol levels, heart rate, tremor intensity, and demographic information.

B. Data Preprocessing Techniques To ensure data consistency and quality, the following preprocessing steps were applied:

1. **Handling Missing Values**:
   - Any missing values were addressed using mean, median, or mode imputation techniques.
   - For categorical data, missing values were filled using the most frequent category.
2. **Data Normalization and Scaling**:
   - Continuous variables such as glucose levels, blood pressure, and cholesterol levels were normalized using Min-Max Scaling.
   - Standardization was applied to ensure uniformity across different feature ranges.
3. **Feature Selection**:
   - Correlation analysis was performed to identify relevant features.
   - Principal Component Analysis (PCA) was applied to reduce dimensionality while preserving essential information.
4. **Data Splitting**:
   - The dataset was divided into training (80%) and testing (20%) subsets.

    - Cross-validation techniques were employed to enhance model robustness.
5. **Class Imbalance Handling**:
    - Synthetic Minority Over-sampling Technique (SMOTE) was applied to balance the dataset.
    - Under-sampling methods were also considered where necessary.
6. **Data Encoding**:
    - Categorical variables such as gender and medical history were encoded using one-hot encoding and label encoding methods.

C. Model Implementation and Evaluation The preprocessed data was fed into multiple machine learning algorithms, including:

- **Logistic Regression**
- **Support Vector Machines (SVM)**
- **Random Forest Classifier**
- **Naïve Bayes Algorithm**
- **K-Nearest Neighbors (KNN)**

Each model was evaluated using the following performance metrics:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-score**
- **ROC-AUC Curve**