Multiple Disease Prediction System

Prepared by

Abhay Suresh Patil (CWID: 885205807)

Submitted to

Dr. Abdul Motin Howlader

Department of Computer Science

California State University, Fullerton

CALIFORNIA STATE UNIVERSITY
FULLERTON™

Department of Computer Science

This project has been satisfactorily demonstrated and is of suitable form.

This project report is acceptable in partial completion of the requirements for the Master of Science degree in Computer Science.

Multiple Disease Prediction System

Project Title (type)

Abhay Suresh Patil

Student Name (type)

Dr. Abdul Motin Howlader

Advisor's Name (type)

Advisor's signature Date

Reviewer's name

Reviewer's signature Date

Abstract

The persistent enlargement in healthcare data and rising complicatedness in interpreting critical conditions have emphasized the importance of designing advanced predictive solutions. Traditional medical diagnostic procedures often suffer from delayed detection, opinionated scalability, and reliance on subjective clinical interpretations. To handle these challenges, this project introduces a Multiple Disease Prediction System (MDPS) that incorporates machine learning algorithms to predict the possibility of Diabetes, Heart Disease, and Parkinson's Disease from clinical datasets. The presented system employs supervised learning algorithms—Support Vector Machines (SVM) for Diabetes and Parkinson's Disease and Logistic Regression for Heart Disease—to classify and predict patient conditions based on relevant clinical and physiological characteristics.

The process involves extensive data collection from standardized medical datasets and stringent preprocessing, feature engineering, and selection to enhance predictive precision. The models were considered using standard metrics, including accuracy, precision, recall, and F1-score, establishing strong and reliable predictive capabilities suitable for real-world medical applications. Also, an interactive, user-friendly predictive interface was designed using the Streamlit framework, enabling healthcare experts and end-users to obtain immediate diagnostic understandings.

The experimental results demonstrate that the designed system simplifies diagnostic processes and improves early-stage detection accuracy, facilitating timely clinical interventions. This project research paper also examines possible integration into clinical practices, restrictions such

as dataset generalization, and future possibilities like developing the prediction to further diseases and deploying the design onto cloud-based healthcare infrastructures.

Keywords: Multiple Disease Prediction, Supervised Learning, Support Vector Machine, Logistic Regression, Healthcare Analytics, Streamlit, Machine Learning Applications.

1. Introduction

1.1 Background

The healthcare industry is experiencing a significant mutation fueled by fast technological improvements, the digitalization of patient records, and the exponential development of medical data. As chronic conditions persist and grow worldwide, traditional diagnostic practices—especially reliant on symptomatic evaluations, clinical testing, and manual interpretation—are increasingly inadequate. Due to late-stage detection and suboptimal treatment strategies, these traditional practices often donate to deferred diagnoses, inconsistent accuracy, and elevated healthcare costs.

Machine Learning (ML), a core artificial intelligence part, has arisen as a promising solution to these challenges. ML methods present diagnostic capabilities that often surpass conventional methods by allowing automated analysis, pattern recognition, and predictive modeling. These algorithms are trained to explore extensive medical data, uncovering subtle patterns and predicting disease outcomes with high precision. Chronic diseases such as Diabetes, Cardiovascular Disease, and Parkinson's Disease—distinguished by complex, multifactorial etiologies and growing global incidence—are especially well-suited for ML-based interventions.

Correct and early diagnosis of such conditions can direct timely treatment, improve clinical outcomes, enhance the quality of life for patients, and reduce long-term healthcare expenses. The current healthcare system produces a wealth of clinical and physiological data, including laboratory test results, medical imaging, genomic information, and data from wearable fitness devices. Despite the quantity of such data, healthcare professionals often face difficulties in entirely using it due to its massiveness, complexity, and variability. In this context, machine learning is crucial in converting raw healthcare data into meaningful insights. ML allows early disease detection through sophisticated predictive modeling, helps the development of personalized medicine regimens, and enables data-driven, evidence-based clinical decision-making.

1.2 Motivation

While ML algorithms have increasingly been embraced in healthcare studies, multiple existing predictive systems have regulations restricting their practical clinical applications. Most predictive measures are developed slightly for single disease detection, resulting in fragmented resolutions that problematize clinical performance and adoption. Such methods often operate independently, causing healthcare doctors to employ multiple platforms for various disease predictions, which adds complexity, uncertainties, and possible inaccuracies in diagnostics. Also, existing predictive frameworks often lack instinctive, user-friendly interfaces, making them inaccessible to clinicians not well-versed in data science or computational techniques. Consequently, healthcare experts might underutilize strong predictive capabilities due to complexity or unsatisfactory system design.

A prominent gap exists between research-stage predictive precision and real-world clinical usability. High-performing predictive algorithms showcased in study settings often fail to solve effectively in routine clinical practice due to various challenges, including model interpretability problems, inefficient integration with hospital approaches, poor robustness to various real-world data, and the absence of efficient real-time interaction capabilities. Therefore, many predictive systems do not gain widespread clinical adoption or meaningful, valuable impact despite their considerable potential.

Inspired by these important gaps and imperfections, this project aims to develop a complete, integrative solution that simultaneously manages multiple disease states within a suitable predictive framework. This approach facilitates diagnostic processes, reduces complexity, and enhances practical usability for medical professionals. By developing a technically robust and practically intuitive system—through progressive machine learning methods and an affordable, real-time predictive interface—this project aims to bridge existing research-practice gaps, making predictive analytics an essential part of everyday healthcare practices.

1.3 Objectives and Contribution

The preliminary objectives of this study include:

- Creating a robust, incorporated predictive system using machine learning algorithms for parallely predicting Diabetes, Heart Disease, and Parkinson's Disease.
- Using supervised machine learning methodologies, Support Vector Machines (SVM) for Diabetes and Parkinson's Disease and Logistic Regression for Heart Disease were selected for their excellent interpretation in clinical data classification schemes.

- Creating a rigorous data pipeline applying data collection, preprocessing, and sophisticated feature engineering to optimize predictive accuracy.
- Executing an accessible, web-based interface using the Streamlit framework allows clinicians and end-users to predict disease in real time and enables direct practical usage.

This project contributes:

- Delivering an integrative predictive model capable of simultaneously diagnosing multiple diseases, significantly enhancing its practical clinical utility.
- It offers comprehensive evaluations that show robust and reliable system performance through key metrics such as precision, recall, and F1-score.
- Delivering a practical, intuitive user interface specifically designed for healthcare professionals, reducing barriers to adoption, and enhancing predictive system accessibility for clinical and non-technical users alike.

2. Literature Review

2.1 Machine Learning in Disease Prediction

Machine learning (ML) has evolved into a transformative force in healthcare analytics, delivering profound advances in disease prediction, diagnosis accuracy, and treatment optimization. ML algorithms excel at discovering complicated connections hidden within extensive and diverse healthcare datasets, thus improving the early and correct diagnosis of different chronic conditions. Over the past decade, multiple studies have exhibited the power of

machine learning procedures in forecasting diseases using clinical, genomic, and physiological data, gaining results exceeding traditional diagnostic approaches in speed and accuracy [1].

Support Vector Machines (SVM), Logistic Regression, Decision Trees, Random Forests, Gradient Boosting Machines, and Deep Neural Networks (DNN) have been widely used in predictive healthcare due to their robust classification interpretation. For example, in predicting Diabetes, SVM and Random Forest algorithms have demonstrated high predictive accuracy by analyzing patient glucose levels, BMI, insulin measures, and genetic markers. A study by Kavakiotis et al. [3] studied Diabetes prediction through various ML models, concluding that ensemble models and SVM algorithms always produced excellent accuracy due to their strong generalization abilities and robustness to dataset variability.

Also, heart disease prediction has broadly employed Logistic Regression, Neural Networks, and ensemble approaches, with Logistic Regression being particularly preferred for its interpretability, clarity, and reliable probability-based forecasts. A study by Mohan et al. [5] compared multiple ML algorithms, including Logistic Regression, Random Forest, and Gradient Boosting, figuring that ensemble-based algorithms like Gradient Boosting offer marginally higher accuracy, although Logistic Regression stays clinically principal due to their clarity and ease of clinical understanding.

In neurodegenerative conditions such as Parkinson's Disease, the application of ML has seen significant progress. SVM and Deep Learning-based classifiers are often embraced, processing acoustic signals, motor symptoms data, and imaging-based biomarkers. Studies such as those by

Little et al. [4] and Tsanas et al. [8] have demonstrated the effectiveness of ML in identifying Parkinson's Disease through voice analysis, with Support Vector Machines emerging as a highly reliable classifier due to their ability to handle high-dimensional data efficiently and robustly.

One of the pivotal strengths of ML-based approaches in disease prediction is their capacity to handle complex, multidimensional datasets that traditional statistical methods struggle with. ML models can significantly improve clinical outcomes' prediction accuracy and interpretability by incorporating advanced feature engineering techniques and robust data preprocessing strategies. ML algorithms also facilitate personalized medicine, adapting prediction models to individual patient profiles and enabling healthcare providers to offer tailored treatments and preventive measures.

Despite significant advancements, challenges remain in machine learning applications within healthcare. Data imbalance, insufficient dataset size, and limited model interpretability persist as barriers to wider clinical adoption. Researchers often emphasize the significance of developing answerable and interpretable measures to improve clinician trust, decision-making clarity, and the ongoing growth within this research area [2], [7].

2.2 Existing Systems and Limitations

The current wave of welfare in predictive healthcare has led to the evolution of multiple automated diagnostic systems, mainly focusing on single-disease prediction. Existing machine learning-driven prediction frameworks show considerable promise in delivering precise disease

detection and improving clinical decision-making. However, these systems often exhibit critical limitations hindering their adoption and practical utility within clinical environments.

One general category of existing prediction models is disease-specific diagnostic tools. Systems such as the Diabetes forecast framework proposed by Perveen et al. [6], the heart disease prediction model developed by Mohan et al. [5], and Parkinson's detection through vocal biomarkers by Tsanas et al. [8] are notable examples. These solutions achieve remarkable predictive accuracy by utilizing ML algorithms and specific datasets tailored to their diseases. However, such isolated systems inherently limit usability in clinical practice, as healthcare professionals must deploy multiple platforms independently, complicating the diagnostic workflow and increasing operational complexity.

Moreover, single-disease systems often neglect the clinical reality that multiple chronic conditions often coexist within individual patients, requiring holistic rather than secluded diagnostic approaches. This compartmentalized standpoint decreases practical effectiveness, as clinicians face the added burden of solving numerous distinct outputs from individual prediction systems, potentially compromising the consistency and dependability of clinical interpretations.

Another critical constraint is linked to the accessibility and usability of existing prediction frameworks. Despite achieving technical excellence, many systems are complicated, demanding specialized data science and calculating ability to operate effectively. This sophistication creates significant barriers, restricting the participation and arrangement of healthcare professionals who may lack comprehensive computational training. Studies taught by Elshawi et al. [2] emphasize

that clinicians frequently prefer prediction models with high interpretability, usability, and direct integration into routine practice rather than highly complex models with marginally superior accuracy. Therefore, overly complex or poorly designed user interfaces significantly hinder such systems' practical adoption and clinical relevance.

Also, existing prediction systems frequently face challenges related to interpretability and transparency, particularly those employing sophisticated algorithms like Deep Neural Networks or ensemble methods. Despite their proven accuracy, such "black box" models lack clear interpretability, making clinicians hesitant to trust the outputs, especially in critical decision-making scenarios. Ribeiro et al. [7] emphasize that interpretability is essential in healthcare, enabling clinicians to justify and confidently make patient diagnostic decisions.

Data-related problems remain fundamental limitations, including insufficient dataset size, absence of diverse and representative clinical data, and data inequality, decreasing model robustness and generalizability. Numerous existing predictive models are trained on narrowly defined populations or limited datasets, harshly restricting their predictive trustworthiness when used to broader, more diverse patient populations in real-world scenarios. Researchers always support more extensive, well-curated, diverse datasets to enhance predictive models' robustness and real-world relevance [1].

3. Methodology

3.1 Proposed System Overview

The presented Multiple Disease Prediction System (MDPS) denotes an advanced, unified machine learning-based framework developed explicitly for predicting the possibility of critical medical conditions, including Diabetes, Heart Disease, and Parkinson's Disease. The design is strategically devised to bridge the current voids followed in clinical practice, notably the fragmentation of single-disease diagnostic tools, limited accessibility, and suboptimal interpretability. The MDPS combines multiple predictive models within a single interactive platform to address these issues comprehensively, guaranteeing accurate, real-time, and coexisting multi-disease projection capabilities.

The architecture of the MDPS contains three primary phases:

3.1.1 Data Acquisition and Preprocessing

In the initial phase, high-quality clinical datasets are systematically collected and preprocessed to ensure optimal data quality. These datasets include:

- ***Diabetes Dataset:*** *Including medical parameters such as glucose levels, insulin measures, BMI, pregnancy count, and other physiological features.*
- ***Heart Disease Dataset:*** *Comprising critical clinical indicators such as cholesterol levels, blood pressure, chest pain types, and electrocardiogram measures.*
- ***Parkinson's Dataset:*** *Employing acoustic and speech signal characteristics like jitter, shimmer, vocal frequency variations, and signal processing metrics.*

Data preprocessing implicates advanced techniques, including normalization, outlier detection, and handling missing values, assuring that all datasets preserve consistency, reliability, and robustness for subsequent research.

3.1.2 Predictive Modeling

The second phase involves robust supervised machine-learning algorithms tailored to each disease state. The MDPS employs the following models:

- ***Support Vector Machine (SVM):*** *Selected for Diabetes and Parkinson's Disease predictions due to their effectiveness in handling high-dimensional datasets and robustness to overfitting, thus securing reliable diagnostic outcomes.*
- ***Logistic Regression:*** *Utilized for Heart Disease prediction because of its interpretability, transparency, and clinical reliability, delivering clinicians with clear and coherent probability-based predictions.*

These predictive measures experience meticulous training, testing, and validation techniques using cross-validation techniques and performance metrics, including accuracy, precision, recall, and F1-score. Hyperparameter tuning and feature selection techniques further enhance model accuracy and efficiency.

3.1.3 Interactive User Interface Development

The last phase commands the result of a real-time predictive interface using the Streamlit framework. This interactive, user-friendly platform allows clinicians and public users to input suitable medical data seamlessly and obtain immediate predictive results. Streamlit is utilized for

its plainness, accessibility, and practical real-time performance abilities, guaranteeing an intuitive user experience and especially lowering barriers to applicable clinical adoption.

3.1.4 System Workflow

The predictive workflow starts with user input collection via the interactive interface, where patient-specific clinical data is documented. Data inputs are processed directly through the individual-trained predictive models upon submission. The developed predictions are immediately envisioned, indicating each disease's presence or absence. This workflow delivers instantaneous and actionable understandings, allowing healthcare professionals to make instructed, timely clinical decisions.

By integrating multiple predictive abilities into a precise, user-focused interface, the MDPS effectively manages significant constraints inherent to conventional single-disease models. This cooperative technique facilitates diagnostic processes and enhances real-world usability and clinical relevance, promising substantial advances in healthcare delivery and patient results.

3.2 Data Collection and Description

Effective disease forecasting depends on robust and factual clinical data, making systematic data collection a crucial element of the suggested predictive framework. To provide high predictive accuracy and model reliability, the Multiple Disease Prediction System (MDPS) datasets were sourced from reputable public medical data repositories, ensuring the data's validity, consistency, and comprehensiveness.

3.2.1 Diabetes Dataset

The diabetes dataset consists of several clinically relevant parameters, capturing essential physiological and biochemical attributes indicative of diabetic conditions. Specifically, the dataset comprises:

- ***Pregnancies:*** *Number of pregnancies (for female patients).*
- ***Glucose Level:*** *Plasma glucose concentration calculated during an oral glucose tolerance test.*
- ***Blood Pressure:*** *Diastolic blood pressure (mm Hg).*
- ***Skin Thickness:*** *Triceps skin fold thickness (mm) indicates subcutaneous fat.*
- ***Insulin Level:*** *2-hour serum insulin measurement (μU/ml).*
- ***Body Mass Index (BMI):*** *Calculated as weight (kg) divided by height squared (m^2).*
- ***Diabetes Pedigree Function:*** *A numerical index capturing genetic predisposition based on family history.*
- ***Age:*** *Age of the patients.*

Each record in the dataset is clearly labeled, indicating diabetic (1) or non-diabetic (0) outcomes. This binary classification facilitates supervised learning algorithms' practical training and evaluation.

3.2.2 Heart Disease Dataset

The heart disease dataset was compiled from the University of California Irvine (UCI) Machine Learning Repository, a significant source of various validated datasets often employed in

machine learning research. The dataset contains key cardiovascular features correlated with heart disease presence. Unique attributes include:

- ***Age and Gender:*** *Fundamental demographic parameters influencing disease prevalence.*
- ***Chest Pain Type (cp):*** *Categorically organized chest pain symptoms (typical Angina, atypical Angina, non-anginal pain, asymptomatic).*
- ***Resting Blood Pressure (treetops):*** *Resting blood pressure (mm Hg) recorded during hospital admission.*
- ***Serum Cholesterol (chol):*** *Serum cholesterol levels (mg/dl).*
- ***Fasting Blood Sugar (fbs):*** *Fasting blood sugar (>120 mg/dl, binary indicator).*
- ***Resting Electrocardiographic Results (resting):*** *ECG results show cardiac electrical movement.*
- ***Maximum Heart Rate (thalach):*** *Highest heart rate reached during a stress test.*
- ***Exercise Induced Angina (exang):*** *Angina triggered by physical exertion.*
- ***ST Depression (old peak):*** *ST-segment depression induced by exercise relative to rest.*
- ***Slope of Peak Exercise ST Segment (slope):*** *Pattern of ST segment during peak exercise.*
- ***Thalassemia (thal):*** *Categorized as a normal, fixed, or reversible defect.*

Like the diabetes dataset, the heart disease dataset includes marked records identifying the presence (1) or absence (0) of heart disease, crucial for supervised predictive modeling.

3.2.3 Parkinson's Disease Dataset

The dataset for Parkinson's Disease was sourced from the UCI Machine Learning Repository, containing acoustic measures broadly used for Parkinson's detection via vocal analysis. This

dataset contains advanced voice and speech-processing features expressive of Parkinsonian conditions, such as:

- ***MDVP (Multi-Dimensional Voice Program) parameters:*** *Fundamental frequency (Fo), maximum frequency (Fhi), minimum frequency (Flo).*
- ***Jitter and Shimmer parameters:*** *Variations in vocal frequency (jitter) and amplitude (shimmer) represent vocal instability.*
- ***Noise-to-Harmonic Ratio (NHR) and Harmonics-to-Noise Ratio (HNR):*** *Metrics quantifying the quality and clarity of voice signals.*
- ***Recurrence Period Density Entropy (RPDE) and Detrended Fluctuation Analysis (DFA):*** *Complex nonlinear vocal metrics capturing subtle voice irregularities.*
- ***Pitch Period Entropy (PPE) and additional frequency modulation parameters:*** *Advanced signal-processing features capturing vocal stability and phonetic deviations.*

Each voice sample record is classified, indicating the presence (1) or absence (0) of Parkinson's Disease.

3.2.4 Data Quality and Ethical Considerations

All datasets used in this research were publicly known, de-identified, and ethically managed, aligning with research standards and privacy regulations. The datasets' demonstrated reliability and prior extensive validation across diverse research studies ensure the robustness of predictive studies performed in this project.

3.3 Data Preprocessing and Feature Engineering

Correct predictive modeling in healthcare analytics depends on the quality and comprehensiveness of the underlying data. Raw clinical datasets often offer numerous challenges, such as inconsistencies, missing values, irrelevant attributes, and data imbalance, potentially degrading the predictive performance of machine learning algorithms. To provide optimal model accuracy and reliability, this investigation employs extensive data preprocessing and sophisticated feature engineering methods to design datasets meticulously before model training.

3.3.1 Data Preprocessing

Data preprocessing symbolizes the foundational step to ensure that datasets are suitable for practical training and predictive analysis. The key preprocessing steps executed in this project include:

- ***Missing Value Treatment:*** *Missing values were determined systematically and managed utilizing statistical imputation methods. Particularly, median imputation was assumed for ongoing numerical features, providing minimal distortion of original data distribution. For categorical or binary variables, mode-based imputation was used, thus keeping necessary categorical distributions within the datasets.*
- ***Outlier Detection and Removal:*** *Medical datasets often have outliers due to dimension errors or genuine clinical variability. Outliers were noticed using statistical methods such as the Z-score and Interquartile Range (IQR) analysis. Identified outliers were handled through reduction (in cases of clearly erroneous values), thus restricting extreme values to decrease negative impacts on model training.*

- ***Data Normalization and Standardization:*** *Normalization and standardization were important preprocessing steps since clinical datasets generally contain numerical elements with varying scales and distributions. Features were mounted using Min-Max normalization, restraining data to a uniform [0,1] range, improving algorithm intersection and prediction stability, particularly for algorithms sensitive to feature scales like SVM. Standardization (Z-score scaling) was used selectively to reach zero-mean and unit-variance distributions where applicable.*

- ***Class Balancing Techniques:*** *Imbalanced types are common in clinical datasets, potentially skewing prediction precision towards majority types. SMOTE (Synthetic Minority Oversampling Technique) and random undersampling methods effectively balanced dataset classes, guaranteeing that machine learning models accurately represent minority-class instances.*

3.3.2 Feature Engineering and Extraction

Feature engineering concerns deriving new, informative features from raw data and improving predictive model effectiveness. This research executed various feature engineering techniques tailored specifically for each dataset:

3.3.2.1 Diabetes Dataset:

- *BMI Categorization: Derived categorical BMI features (Underweight, Normal, Overweight, Obese) to enable BMI interpretation for modeling.*
- *Age Grouping: Ordered age into medically suitable groups to capture age-related diabetes risk.*
- *Glucose-Insulin Ratio: Computed ratios to represent insulin resistance more precisely and enhance prediction.*

3.3.2.2 Heart Disease Dataset:

- *Risk Factor Aggregation: Connected cholesterol, blood pressure, and fasting blood sugar levels into a suitable cardiovascular risk score.*
- *Exercise-Related Features: Assembled new features combining exercise-induced angina, maximum heart rate, and ST-segment slope to capture exertion-related risk factors better.*

3.3.2.3 Parkinson's Disease Dataset:

- *Advanced Voice Metrics: Engineered voice-derived features by connecting jitter and shimmer parameters to effectively capture vocal instability indicative of Parkinsonian conditions.*
- *Signal Processing Metrics: Combined recurrence period density entropy (RPDE), detrended fluctuation analysis (DFA), and pitch period entropy (PPE) into unified non-linear speech features, catching subtle vocal irregularities to enhance model discrimination capabilities.*

3.3.3 Feature Selection

Additionally, feature selection techniques were used to enhance predictive accuracy, reduce overfitting, and improve interpretability. Recursive Feature Elimination (RFE), correlation analysis, and Chi-Square statistical tests were used. Feature selection significantly reduced the dimensionality of datasets, providing that only highly relevant features were included. This dimensionality reduction optimized computational efficiency and enhanced the generalization capabilities of the models.

This research produced robust, high-quality datasets for the subsequent predictive modeling stage through rigorous preprocessing, advanced feature engineering, and effective feature selection. The processed datasets provided the highest predictive potential, precision, and interpretability, applying a solid basis for reliable multiple disease prediction using machine learning algorithms.

3.4 Machine Learning Algorithms

The performance and reliability of a clinical prediction system depend on the choice of machine learning algorithms. In this study, algorithm selection was disease-specific to align with the nature of the data and prediction goals.

3.4.1 Support Vector Machine (SVM)

Applied to: Diabetes and Parkinson's Disease

SVM was selected for its significance in binary classification and its ability to handle high-dimensional, nonlinear datasets. For diabetes prediction, SVM provided a robust dataset

interpretation by distinguishing between diabetic and non-diabetic cases utilizing features like glucose and BMI. In Parkinson's disease prediction, SVM handled acoustic elements well, keeping accuracy despite the small dataset size.

Clinical Strengths: High generalization, reasonable with complex datasets, low risk of overfitting.

3.4.2 Logistic Regression (LR)

Applied to: Heart Disease

LR was selected for its clarity, interpretability, and alignment with clinical risk models. It operates a logistic function to assess disease probability and delivers transparent coefficients describing the influence of clinical characteristics such as age, cholesterol, and blood pressure.

Clinical Strengths: Manageable to interpret, trusted by clinicians, fast to implement.

3.4.3 Model Training and Evaluation

Models were trained using an 80/20 stratified split and validated with 5-fold cross-validation. Grid Search was used to tune key hyperparameters for both SVM and LR. Performance was calculated utilizing accuracy, precision, recall, and F1-score. Final models were serialized using Python's pickle library for integration into the system's web interface.

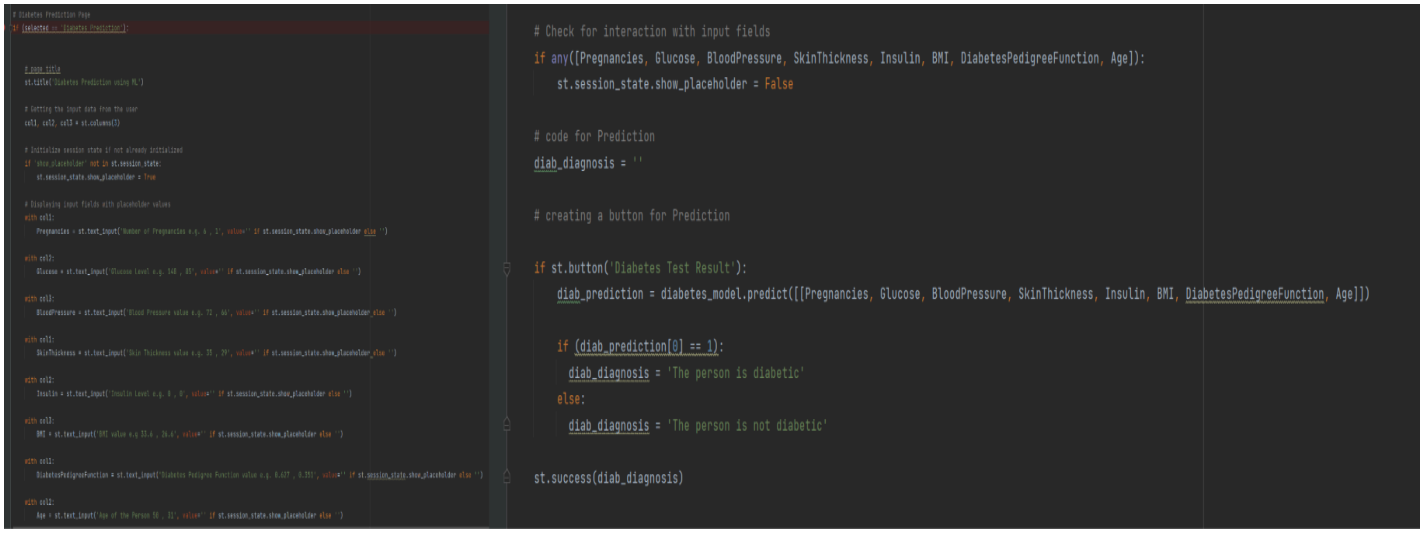
4. Implementation

```python
import pickle
import streamlit as st
from streamlit_option_menu import option_menu


# loading the saved models

diabetes_model = pickle.load(open('C:/Users/Abhay Patil/Multiple-Disease-Prediction-System/saved models/diabetes_model.sav', 'rb'))

heart_disease_model = pickle.load(open('C:/Users/Abhay Patil/Multiple-Disease-Prediction-System/saved models/heart_disease_model.sav','rb'))

parkinsons_model = pickle.load(open('C:/Users/Abhay Patil/Multiple-Disease-Prediction-System/saved models/parkinsons_model.sav', 'rb'))



# sidebar for navigation
with st.sidebar:

    selected = option_menu('personal health guardian',

                           ['Diabetes Prediction',
                            'Heart Disease Prediction',
                            'Parkinsons Prediction'],
                           icons=['activity','heart','person'],
                           default_index=0)
```

Fig. 1. Model Loading and Navigation Setup

The first section of the implementation code consists of setting up the computational environment by integrating essential Python modules and restoring the pre-trained machine-learning models from the computer. Using the pickle library, models earlier trained to predict Diabetes, Heart Disease, and Parkinson's Disease are loaded into memory in serialized .sav formats. This approach ensures that model inference can be executed without extra training overhead. The interface layer is built using Streamlit, a Python framework tailored to develop interactive web applications rapidly. At that exact time, streamlit_option_menu is added to construct an aesthetically arranged, user-friendly sidebar navigation system.

Tracking model initialization, a vertical navigation menu is instantiated within the sidebar to allow seamless changes between three distinct diagnostic interfaces: one for Diabetes, Heart Disease, and Parkinson's Disease. A separate menu entry is linked with a representative icon, which contributes to the intuitiveness and accessibility of the platform. The menu selection dynamically governs which diagnostic module is rendered in the primary workspace, enabling a modular and maintainable code structure. This segmentation technique improves clarity from the user's perspective and facilitates separate development and testing of each disease-specific predictive module.



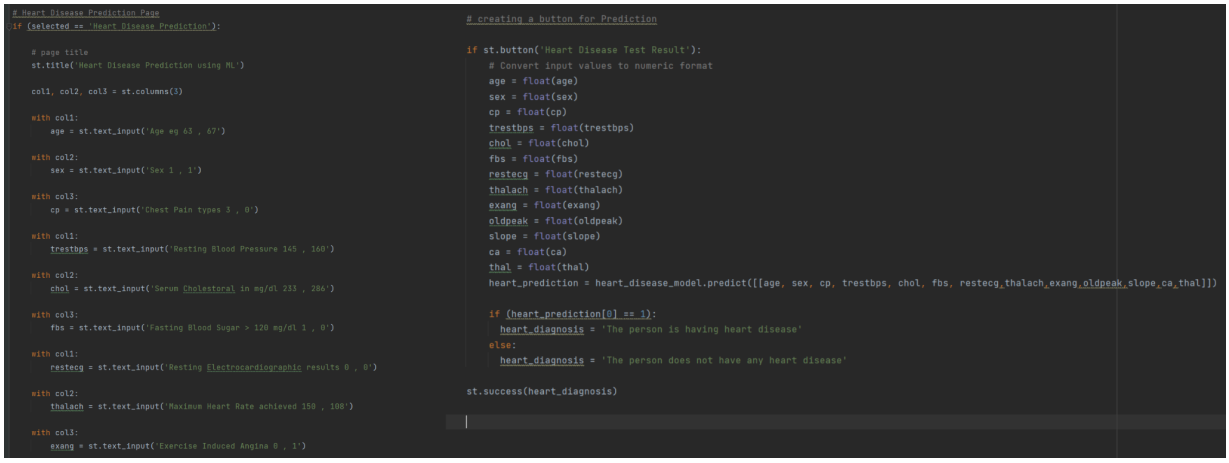
```
# Check for interaction with input fields
if any([Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age]):
    st.session_state.show_placeholder = False

# code for Prediction
diab_diagnosis = ''

# creating a button for Prediction

if st.button('Diabetes Test Result'):
    diab_prediction = diabetes_model.predict([[Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age]])

    if (diab_prediction[0] == 1):
      diab_diagnosis = 'The person is diabetic'
    else:
      diab_diagnosis = 'The person is not diabetic'

st.success(diab_diagnosis)
```

Fig. 2. Diabetes Prediction Module Interface

The second part of the code deals with helping the Diabetes prediction module. When the user selects "Diabetes Prediction" from the sidebar, the project detects this option and dynamically generates a revived section in the main interface titled "Diabetes Prediction using ML." The interface layout uses a three-column format (col1, col2, col3) to provide a methodical and readable format. Before displaying input fields, the system restricts whether a show_placeholder

flag is attending in the session state—a mechanism Streamlit operates to maintain variables across reruns. If the flag is not initialized, it is set to True, which provides that placeholder text (e.g., example values) is observable in each input field for user guidance. This placeholder tool remains active until the user interacts with any of the input areas, upon which the flag is updated to turn off the further display of hints, thereby preserving the interface neat on following interactions.

The module contains eight critical clinical parameters from the user: number of pregnancies, glucose level, blood pressure, skin thickness, insulin level, body mass index (BMI), diabetes pedigree function, and age. These values are stored as strings using Streamlit's st.text_input() method and distributed across columns for clearness. Once the user provides the needed input values and connects the "Diabetes Test Result" button, the gathered data is passed as a list to the predict() method of the pre-loaded SVM-based diabetes model. The model processes the input and yields a binary classification. Based on this prediction, the system develops a human-readable diagnosis message—either demonstrating the presence of diabetes or exhibiting its absence. This announcement is then depicted in real time using st.success(), showing the user prompt feedback within the same interface window. This seamless flow from user input to model inference and effect display provides a spontaneous and efficient diagnostic experience.



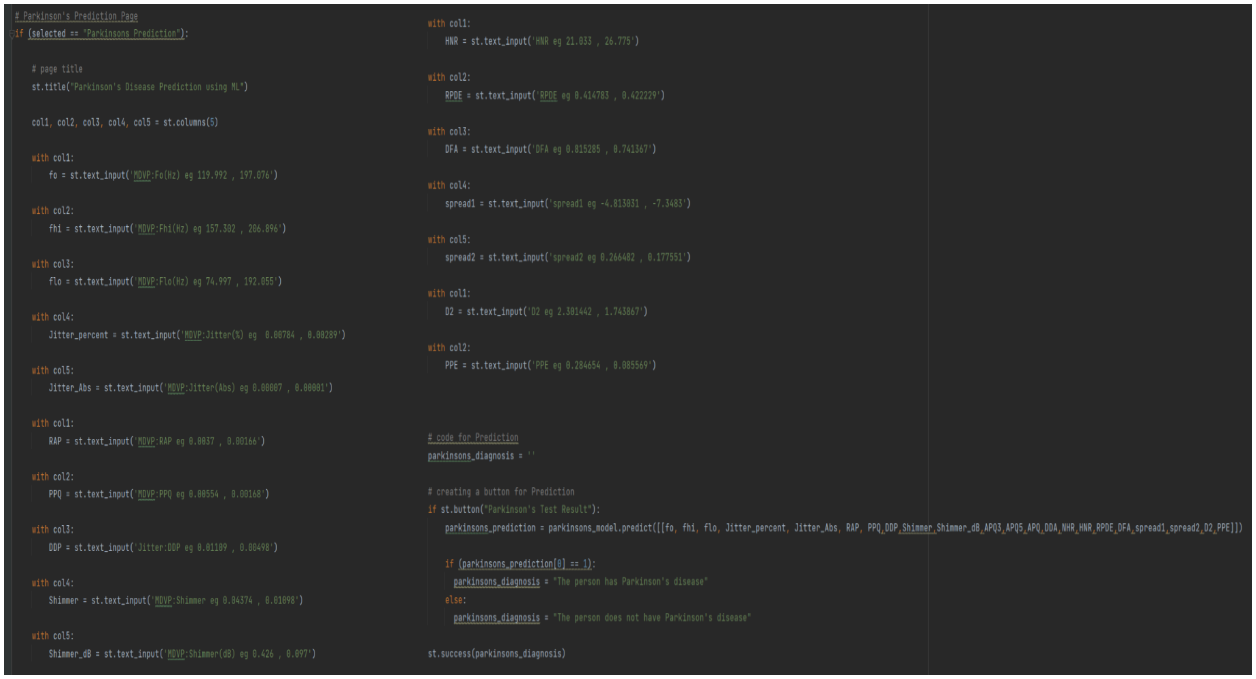
```python
# Heart Disease Prediction Page
if (selected == 'Heart Disease Prediction'):

    # page title
    st.title('Heart Disease Prediction using ML')

    col1, col2, col3 = st.columns(3)

    with col1:
        age = st.text_input('Age eg 63 , 67')

    with col2:
        sex = st.text_input('Sex 1 , 1')

    with col3:
        cp = st.text_input('Chest Pain types 3 , 0')

    with col1:
        trestbps = st.text_input('Resting Blood Pressure 145 , 160')

    with col2:
        chol = st.text_input('Serum Cholestoral in mg/dl 233 , 286')

    with col3:
        fbs = st.text_input('Fasting Blood Sugar > 120 mg/dl 1 , 0')

    with col1:
        resteg = st.text_input('Resting Electrocardiographic results 0 , 0')

    with col2:
        thalach = st.text_input('Maximum Heart Rate achieved 150 , 108')

    with col3:
        exang = st.text_input('Exercise Induced Angina 0 , 1')

    # creating a button for Prediction

    if st.button('Heart Disease Test Result'):
        # Convert input values to numeric format
        age = float(age)
        sex = float(sex)
        cp = float(cp)
        trestbps = float(trestbps)
        chol = float(chol)
        fbs = float(fbs)
        restecg = float(restecg)
        thalach = float(thalach)
        exang = float(exang)
        oldpeak = float(oldpeak)
        slope = float(slope)
        ca = float(ca)
        thal = float(thal)
        heart_prediction = heart_disease_model.predict([[age, sex, cp, trestbps, chol, fbs, restecg,thalach,exang,oldpeak,slope,ca,thal]])

        if (heart_prediction[0] == 1):
          heart_diagnosis = 'The person is having heart disease'
        else:
          heart_diagnosis = 'The person does not have any heart disease'

    st.success(heart_diagnosis)
```

Fig. 3. Heart Disease Prediction Module Interface

Upon choosing "Heart Disease Prediction" from the sidebar, the project will automatically render a new interface developed to estimate the probability of heart disease based on clinical parameters. The user interface starts with a defining page title and embraces a three-column layout to categorize input fields effectively across the screen. Each column captures some of the needed patient data, improving usability and form readability. The module prompts the user to join thirteen medically appropriate characteristics such as age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression, slope of the ST segment, number of major vessels, and the Thalassemia condition.

These elements are assembled using text input fields, each distinctly marked with example values for more helpful user guidance. The user starts the forecast once the data is documented by connecting the "Heart Disease Test Result" button. Before the prediction is accomplished, the input strings are explicitly transformed into floating-point numerical values to provide compatibility with the model's desired input format. These values are then given to a pre-trained

Logistic Regression model that considers the intake and produces a binary classification—suggesting either the existence or absence of heart disease. The outcome is analyzed and transformed into a user-friendly diagnosis message, which is then shown interactively within the application interface. This flow—from data entry to real-time prediction—enables a streamlined and clinically significant diagnostic understanding for the user.



```
# Parkinson's Prediction Page
if (selected == "Parkinsons Prediction"):

    # page title
    st.title("Parkinson's Disease Prediction using ML")

    col1, col2, col3, col4, col5 = st.columns(5)

    with col1:
        fo = st.text_input('MDVP:Fo(Hz) eg 119.992 , 197.076')

    with col2:
        fhi = st.text_input('MDVP:Fhi(Hz) eg 157.302 , 206.896')

    with col3:
        flo = st.text_input('MDVP:Flo(Hz) eg 74.997 , 192.055')

    with col4:
        Jitter_percent = st.text_input('MDVP:Jitter(%) eg  0.00784 , 0.00289')

    with col5:
        Jitter_Abs = st.text_input('MDVP:Jitter(Abs) eg 0.00007 , 0.00001')

    with col1:
        RAP = st.text_input('MDVP:RAP eg 0.0037 , 0.00166')

    with col2:
        PPQ = st.text_input('MDVP:PPQ eg 0.00554 , 0.00168')

    with col3:
        DDP = st.text_input('Jitter:DDP eg 0.01109 , 0.00498')

    with col4:
        Shimmer = st.text_input('MDVP:Shimmer eg 0.04374 , 0.01098')

    with col5:
        Shimmer_dB = st.text_input('MDVP:Shimmer(dB) eg 0.426 , 0.097')

    with col1:
        HNR = st.text_input('HNR eg 21.033 , 26.775')

    with col2:
        RPDE = st.text_input('RPDE eg 0.414783 , 0.422229')

    with col3:
        DFA = st.text_input('DFA eg 0.815285 , 0.741367')

    with col4:
        spread1 = st.text_input('spread1 eg -4.813031 , -7.3483')

    with col5:
        spread2 = st.text_input('spread2 eg 0.266482 , 0.177551')

    with col1:
        D2 = st.text_input('D2 eg 2.301442 , 1.743867')

    with col2:
        PPE = st.text_input('PPE eg 0.284654 , 0.085569')



    # code for Prediction
    parkinsons_diagnosis = ''

    # creating a button for Prediction
    if st.button("Parkinson's Test Result"):
        parkinsons_prediction = parkinsons_model.predict([[fo, fhi, flo, Jitter_percent, Jitter_Abs, RAP, PPQ,DDP,Shimmer,Shimmer_dB,APQ3,APQ5,APQ,DDA,NHR,HNR,RPDE,DFA,spread1,spread2,D2,PPE]])

        if (parkinsons_prediction[0] == 1):
          parkinsons_diagnosis = "The person has Parkinson's disease"
        else:
          parkinsons_diagnosis = "The person does not have Parkinson's disease"

    st.success(parkinsons_diagnosis)
```

Fig. 4. Parkinson's Disease Prediction Module Interface

The last module in the project framework handles the prediction of Parkinson's Disease. When a user selects the "Parkinson's Prediction" option from the sidebar, the tool renders a technical interface labeled "Parkinson's Disease Prediction using ML." This interface is organized into five columns to adjust the extensive number of biomedical voice features required for the diagnosis. The user is prompted to input twenty-two extra parameters derived from sustained vowel

phonation measures and nonlinear signal processing metrics. These contain frequency-based metrics (such as MDVP: Fo, MDVP: Fhi, MDVP: Flo), jitter and shimmer values, harmonic-to-noise ratios, and dynamic features like RPDE, DFA, spread measures, and recurrence parameters.

Individually, the input field is accompanied by a model to guide the user in documenting accurate values. These components were selected based on their clinical applicability in determining neuromotor impairment patterns commonly associated with Parkinson's Disease. Once all values are documented, the user can trigger the forecast by clicking the "Parkinson's Test Result" button. Internally, the input data is collected and handed to a pre-trained Support Vector Machine (SVM) standard tailored for this classification task. The model analyzes the input and outputs a binary result displaying whether Parkinson's Disease likely impacts the person. The development is analyzed and portrayed in real-time via a success notification, either affirming or negating the existence of the disease. This module's strategy balances technical complexity with usability, allowing high-dimensional diagnostic modeling in an intuitive and convenient format.

5. Results

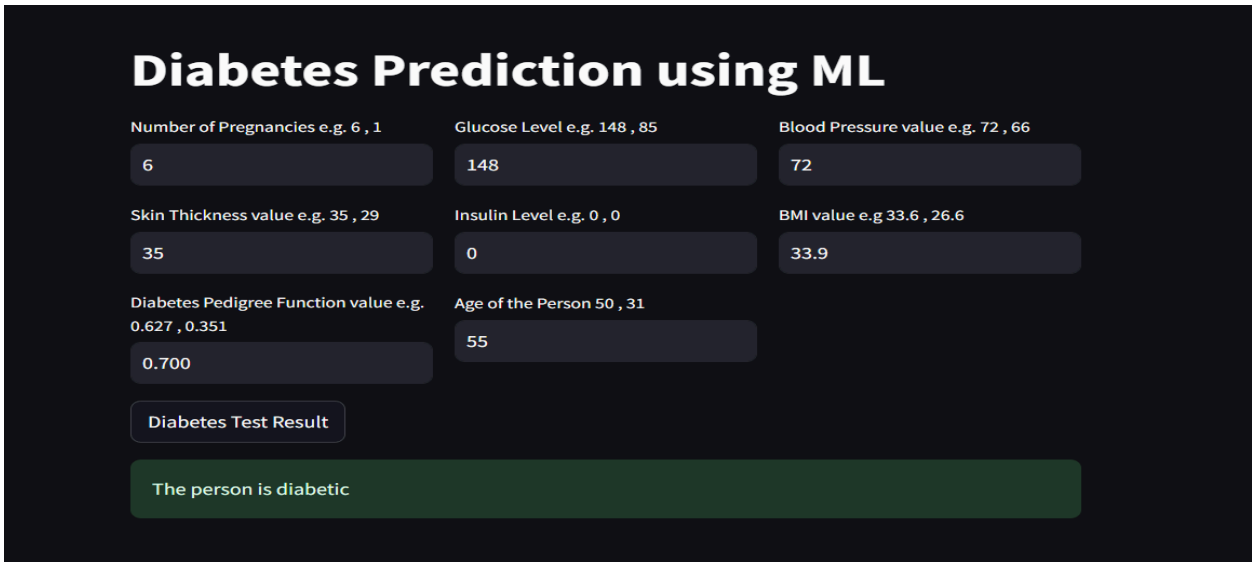


Fig. 5. Diabetes Prediction Output Interface

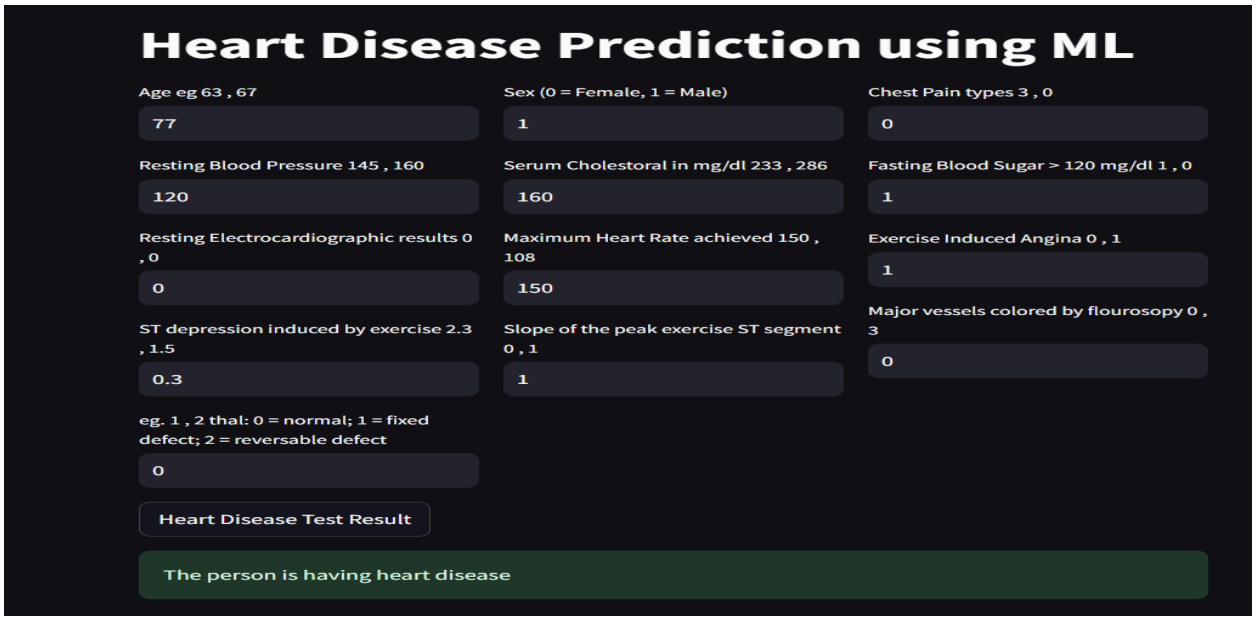


Fig. 6. Heart Disease Prediction Output Interface

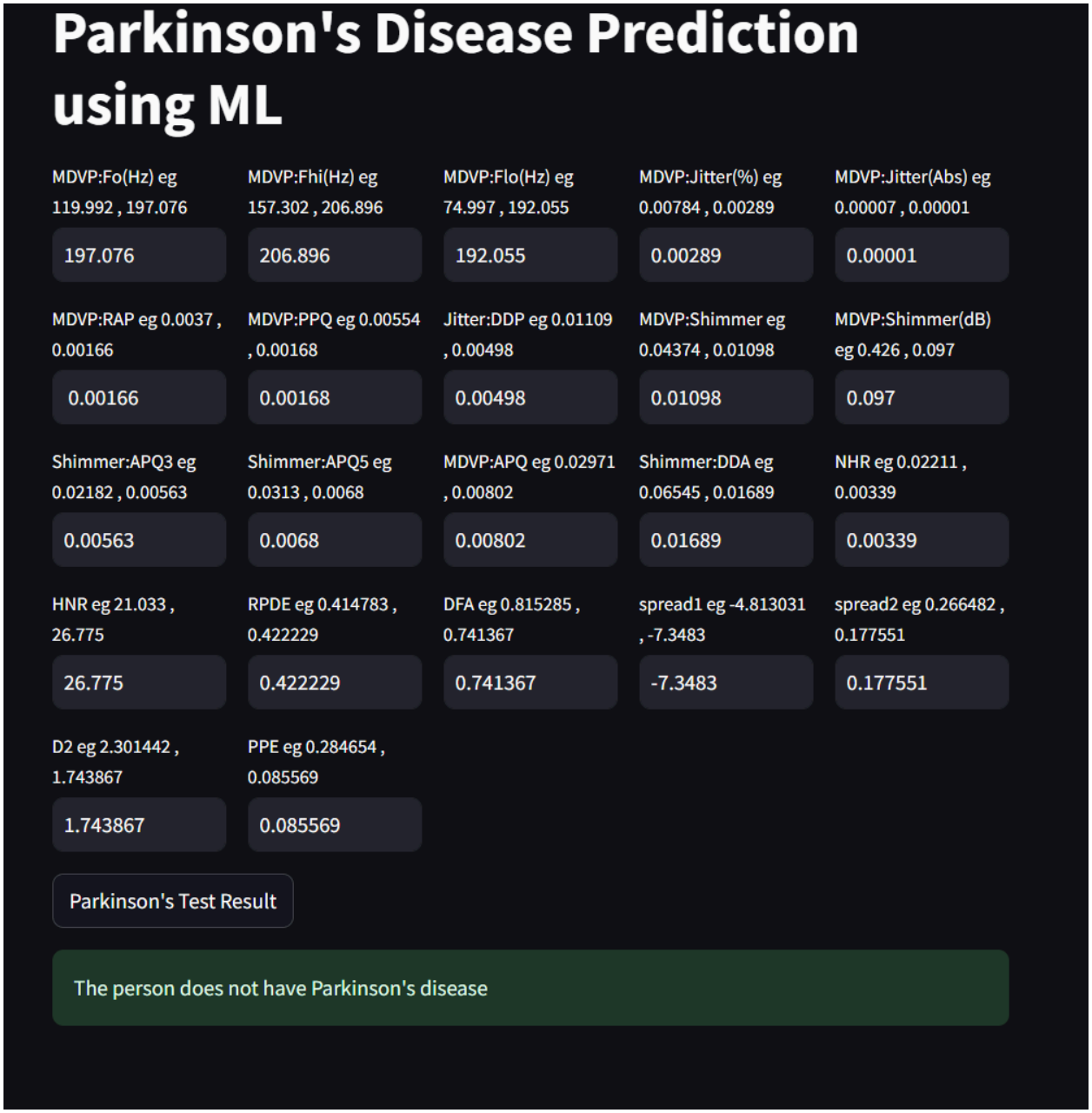


Fig. 7. Parkinson's Disease Prediction Output Interface

6. Future Scope

- Real-time connection with patient electronic health records (EHRs) can help dynamic, up-to-date projections, improving clinical decision-support abilities.
- The project can be extended using domain-specific datasets to forecast other chronic and acute conditions such as liver disease, cancer, and neurological diseases.
- Deploying the project work on cloud platforms like AWS, Microsoft Azure, or Google Cloud will allow scalable access and aid multi-platform interoperability.
- Forthcoming interpretations may utilize advanced deep learning architectures such as convolutional neural networks (CNNs) and long short-term memory (LSTM) networks to capture better-complicated patterns in heterogeneous medical data, including time series and imaging modalities.
- Creating a mobile-compatible version can help remote monitoring and rapid screening, which is especially advantageous in underserved or rural regions.
- The project can deliver constant monitoring and proactive health alerts by interfacing with real-time data from wearable fitness trackers like Fitbit/ Apple Watch.
- Sustaining regional languages and voice-driven interfaces can enhance accessibility and usability for a wider demographic.
- Integration of explainability methods such as SHAP or LIME will enhance confidence and transparency by elucidating model reasoning behind predictions.
- Utilizing federated learning techniques can help measure activity across decentralized data sources while maintaining patient privacy and guaranteeing regulatory compliance with frameworks like HIPAA and GDPR.

- Partnerships with healthcare organizations for clinical trials and verification analyses will be important for achieving regulatory licenses and providing reasonable adoption in medical environments.

7. Conclusion

This research presents a unified and intelligent approach toward early disease prediction by integrating machine learning models into a single, accessible system. Unlike isolated diagnostic tools, the Multiple Disease Prediction System can analyze structured input data to evaluate the risk of three distinct medical conditions—Diabetes, Heart Disease, and Parkinson's Disease. The strength of this work lies not only in the accuracy of its predictions but also in the seamless integration of data science with healthcare usability.

By selecting algorithms, constant performance across diverse datasets, and creating an intuitive user interface, this system contemplates what modern predictive medicine could look like—efficient, proactive, and user-focused. It also applies a solid foundation for future work, from developing disease coverage to integrating real-time data from wearable devices and electronic health records.

While many sophisticated diagnostic tools are being researched today, this project stands out because of its commitment to simplicity without compromising technical integrity. It is designed to be deployed, scaled, and used—not just studied. That shift from theoretical to practical is what gives this work lasting relevance.

This system drives us one step closer to personalized, data-driven healthcare. It indicates that with the right combination of models, data, and human-centered design, technology can bridge medical access gaps and empower clinicians and patients.

It does not merely represent a functional tool but a glimpse into a future where data empowers diagnosis and where technology and humanity move in harmony toward better care.

References

[1] Chen, I. Y., Joshi, S., Ghassemi, M., & Li, M. (2022). Challenges for machine learning in clinical translation of big data. Neuron, 110(9), 1448–1464. https://doi.org/10.1016/j.neuron.2022.03.014

[2] Elshawi, R., Sherif, Y., Al-Mallah, M. H., & Sakr, S. (2019). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. Computational Intelligence, 37(4), 1633–1650. https://doi.org/10.1111/coin.12410

[3] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

[4] Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Transactions on Biomedical Engineering, 56(4), 1015–1022. https://doi.org/10.1109/TBME.2008.2005954

[5] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. International Journal of Recent Technology and Engineering, 8(1), 944–950.

https://www.researchgate.net/publication/333888974_Effective_Heart_Disease_Prediction_Using_Hybrid_Machine_Learning_Techniques

[6] Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science, 82, 115–121. https://doi.org/10.1016/j.procs.2016.04.016

[7] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. arXiv Preprint. https://arxiv.org/abs/1602.04938

[8] Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. IEEE Transactions on Biomedical Engineering, 57(4), 884–893. https://doi.org/10.1109/TBME.2009.2036000

[9] OpenAI. (2023, March 14).[Large language model] https://chat.openai.com