

Prediction of Diabetes, Cardiac and Parkinson's disease using machine learning and streamlit framework

Sumeet Das¹ · Gireesh Kumar¹ · Richa Sharma²

Received: 27 October 2023 / Revised: 10 July 2024 / Accepted: 5 September 2024

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

The development of new scientific instruments and the emergence of new technologies have greatly streamlined, expedited, and eased human life. However, just like a coin has two sides, the advent of development has drawbacks of its own. Specifically leading to a more inactive and sedentary as well as numerous lifestyle diseases like diabetes and elevated stress levels, all of which contribute to a host of cardiac diseases. Extended use of social media devices damages the retina and affects neurons, which can result in Parkinson's disease. The aim of this research is to forecast the probability of lifestyle diseases by utilizing diverse Neural Networks and Machine Learning Algorithms. The algorithms that are employed are Random Forest Classifier, XGBoost, and SVM. Moreover, this prediction system has an interactive user interface that can be deployed and hosted either locally or on world wide web and can be accessed by all.

Keywords Diabetes · Cardiac Disease · Parkinson's Disease · Neural Networks · Machine Learning · Support Vector Machine · XGBoost

1 Introduction

Disease is any harmful deviation or abnormality that affects the structure and function of all or some part of the organism. Disease can broadly be classified into communicable and non-communicable diseases. Lifestyle diseases are the part of the non-communicable diseases. These are the disease caused due to prolonged exposure to modifiable lifestyle behavior like unhealthy eating habits and physical inactivity. In the modern world, where technology is advancing swiftly, various tools have been created that have made human life

✉ Gireesh Kumar
gireesh8@gmail.com; gireesh.kumar@jaipur.manipal.edu

Sumeet Das
sumeetdas28@gmail.com

Richa Sharma
aligarh.richa@gmail.com

¹ Department of CSE, Manipal University Jaipur, Jaipur 303007, India

² Department of Mathematics, JK Lakshmipat University, Jaipur 302026, India

more comfortable and sedentary. However, advancement and development have their own drawbacks, such as an increase in sedentary and inactive way of living as well as several lifestyle diseases like diabetes and a rise in stress levels both of which contribute to various cardiac diseases.

Diabetes, Cardiac and Parkinson's diseases are three of the most prevalent chronic illnesses globally, affecting millions of people every year. Supplementing it are the various fast foods, unruly lifestyles, stress, and mental upheaval has given rise to these various lifestyle diseases [1, 2]. Over the last few years, the lifestyle diseases as mentioned in Fig. 1 have been seen to take a steep increase in a passive form of lifestyle. According to research conducted by the Institute for Health Metrics and Evaluation, it has been seen that the death rate for the patients suffering with type-1 (the insulin-dependent diabetes mellitus or juvenile diabetes) and type-2 (the adult-onset diabetes) has increased with time along with increase in the number of patients with DALY's i.e., Disability-Adjusted Life Years due to accompanying diseases like kidney failure, reduced blood flow and increase in risk of heart attacks in adults [3].

Another more prominent and equally dangerous disease is cardiovascular disease. Due to increase in dependence of technology, hypertension levels and obesity along with growth in tobacco consumption among young adults has altered the facts that adults were mostly affected by heart attacks. and now according to an estimate 1 in 5 heart attacks patients are less than 40 years old and early prediction and treatment is need of the hour [4]. Moreover, another disease discussed under the scope of this paper is the Parkinson's disease is a brain ailment that results in unintentional and uncontrolled movements as well as stiffness, balance issues, and coordination problems. Early detection and cure are extremely important the rate of worsening of the disease is exponential and the patient suffering from it can have severe mental and behavioral changes and difficulty in walking and taking [5, 6].

Early detection and prediction of these diseases can lead to effective treatment and management, reducing the risk of severe problems and improving the quality of life for patients.

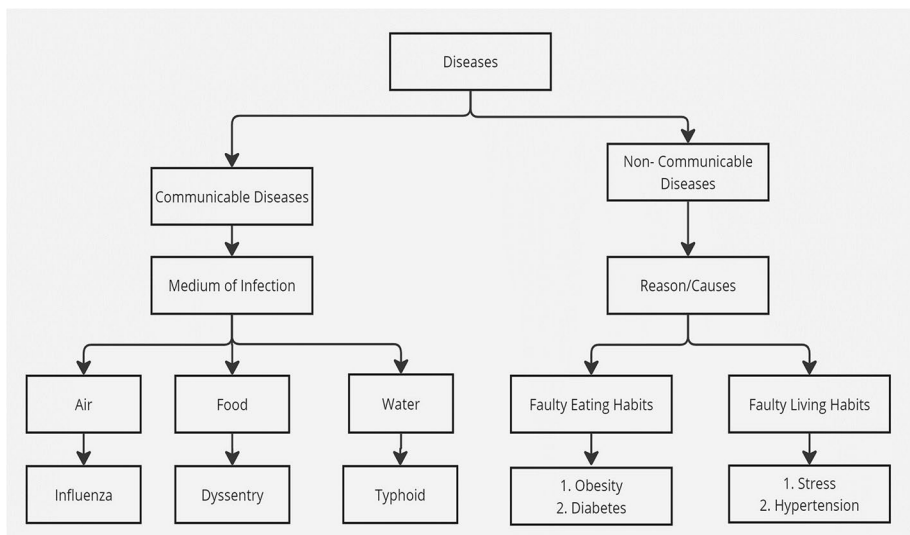


Fig. 1 Classification of Diseases

Machine learning algorithms have shown promise in predicting the onset of these diseases, using various types of data such as medical history, physical exams, and demographic information. In recent years, there has been a significant increase in research related to the use of machine learning algorithms for the prediction of these diseases. Using machine learning techniques, Alic et al. classified diabetes and cardiovascular diseases [7]. They gave an overview of artificial neural networks (ANNs) and Bayesian networks (BNs) used in machine learning to categorise CVDs and diabetes. Yaganteeswarudu developed multi disease prediction model using Machine Learning and Flask API. A system has been proposed that makes use of the Flask API to forecast various diseases. Moreover, analyses of diabetes, diabetes retinopathy, heart disease, and breast cancer were conducted [8]. Using machine learning techniques, Krishnamoorthi et al. investigated a novel diabetes healthcare disease prediction framework [9]. Dhanka and Maini used random forest to identify cases of heart disease [10]. Using machine learning, Abdellatif et al. developed a model for detecting heart disease and classifying it according to severity [11]. Chaki et al. provided a review on diabetes mellitus detection using machine learning and artificial intelligence [12]. Arumugam et al. worked on prediction on multiple disease using Machine learning [13]. Gupta et al. worked on image classification using deep convolutional neural network [14, 15]. For predicting the co-occurrence of diabetes and cardiovascular diseases, Abdalrada et al. developed a machine learning model [16].

This study aims to explore and analyze the current state of the art in the field of predicting Diabetes, Cardiac and Parkinson's disease using machine learning algorithms. This study motivates due to the increment of the number of cases for said diseases. With advent of the pandemic a lot of work has shifted to online leading to a more sedentary life. In this section, the existing methods of disease predictions are reviewed with their benefits and their limitations to emphasize the need for development of a new model. The goal of this paper is to propose and implement web software that can receive input from the user and use algorithms such as SVM, XGBooster, and Random Forest Classifier to predict with reasonable accuracy whether the user has any of the diseases mentioned above. The analysis also discusses the accuracy levels attained by the different algorithms and their performance metrics. The organization of the paper is given as follows. Dataset description has been provided in Section 2. Further, model description is given in Section 3. Results are discussed in Section 4. Finally, concluding observation is specified in Section 5.

2 Dataset description

In this paper, three data sets have been used for predicting the Diabetes (Dataset 1), Parkinson's disease (Dataset 2) and Heart Disease (Dataset 3) respectively. Further, the cumulative description about structure of the datasets used under the scope of this study and the measures calculated after the removal of null and redundant values are provided in Tables 1, 2, 3 and 4 which provides clarity in understanding the diseases subjected to study and analyze.

3 Model description and integration

A unified platform for predicting diabetes, heart disease, and Parkinson's disease has been developed using Streamlit to integrate these models into a single interactive web application.

Table 1 Cumulative description of the datasets

Dataset Parameter	Dataset 1	Dataset 2	Dataset 3
Number of instances	768	195	303
Number of Attributes	8	24	14
Number of Target Variable	1	1	1
Target Variable	A binary variable indicating whether a patient has diabetes or not	A binary variable indicating the health status of the subject	A binary variable indicating whether a patient has heart disease or not
Type of Data	Tabular Data	Tabular Data	Tabular Data

Table 2 Measures obtained after dataset cleaning for diabetes prediction

	Pregnacies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Table 3 Measures obtained after dataset cleaning for heart disease prediction

	age	sex	cp	trestbps	chol	fbis	restecg
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000
	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Table 4 Measures obtained after dataset cleaning for parkinson's disease prediction

	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Fitter(%)	MDVP:Fiter(Abs)	MGVP:RAP	MGVP:PPQ	Jitter:DDP	MDVP:Shimmer
count	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000
mean	154.228641	197.104918	116.324631	0.006220	0.000044	0.003306	0.003446	0.009920	0.029709
std	41.390065	91.491548	43.521413	0.004848	0.000035	0.002968	0.002759	0.008903	0.018857
min	88.333000	102.145000	65.476000	0.001680	0.000007	0.000680	0.000920	0.002040	0.009540
25%	117.572000	134.862500	84.291000	0.003460	0.000020	0.001660	0.001860	0.004985	0.016505
50%	148.790000	175.829000	104.315000	0.004940	0.000030	0.002500	0.002690	0.007490	0.022970
75%	182.769000	224.205500	140.018500	0.007365	0.000060	0.003835	0.003955	0.011505	0.037885
max	260.105000	592.030000	239.170000	0.033160	0.000260	0.021440	0.019580	0.064330	0.119080
	MDVP:Shimmer(dB)	...	Shimmer:DDA	NHR	HNR	status	RPDE	DFA	spreadI
count	195.000000	...	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000
mean	0.282251	...	0.046993	0.024847	21.885974	0.753846	0.498536	0.718099	-5.684397
std	0.194877	...	0.030459	0.040418	4.425764	0.431878	0.103942	0.055336	1.090208
min	0.085000	...	0.013640	0.000650	8.441000	0.000000	0.256570	0.574282	-7.964984
25%	0.148500	...	0.024735	0.005925	19.198000	1.000000	0.421306	0.674758	-6.450096
50%	0.221000	...	0.038360	0.011660	22.085000	1.000000	0.495594	0.722254	-5.720868
75%	0.350000	...	0.060795	0.025640	25.075500	1.000000	0.587562	0.761881	-5.046192
max	1.302000	...	0.169420	0.314820	33.047000	1.000000	0.685151	0.825288	-2.434031

Diabetes prediction: Utilize a Gaussian Naive Bayes classifier, MLP classifier and SVM to predict diabetes based on input features.

Heart disease prediction: Implement logistic regression and an XGB classifier to assess the likelihood of heart disease.

Parkinson's disease prediction: Employ a Random Forest classifier and a neural network model with specified dense layers for Parkinson's disease prediction.

Streamlit integration: Combine all these models into a cohesive web app using Streamlit, allowing users to input relevant health data and receive predictions for each condition through a user-friendly interface. Streamlit facilitates real-time data input, model execution, and display of results, providing a seamless user experience.

Multi-class multi-disease prediction model has been created using SVM, XGBoost, Random Forest classifier and hybridizing the various models to get a better accuracy and output. Brief introduction about all these machine learning algorithms is given below.

In the case for diabetes prediction, three different models have been used namely Gaussian Naïve Bayes, Gaussian distribution, and Multilayer perceptron. In Gaussian Naïve Bayes accounts for feature independence and Gaussian distribution for continuous variables. It is a simple and probabilistic framework and is effective with the dataset used under this study. Moreover, Multilayer perceptron is a neural network with multiple layers and non-linear activation function. It is flexible, powerful and can capture complex patterns. Another model that is used for the diabetes dataset is the Support Vector Machine which finds the optimal hyperplane that helps in maximizing the margin between classes and has robust performance.

Moving on to heart disease prediction, two models have been used namely Logistic Regression and Extreme Gradient Boosting. Logistic Regression decides a linear boundary for binary classification and is simple to implement, effective for binary tasks. The next model used is Extreme Gradient Boosting that ensemble decision trees through boosting and has a high accuracy and is capturing complex patterns along with having a robust performance.

Lastly, under Parkinson's disease prediction Random Forest Model has been used which constructs multiple decision trees and outputs the mode of their prediction robustly and accurately. Also, Neural Network has been implemented having the architecture of input layer with 11 dimensions, dense layers with 8 and 14 units using ReLU activation function and a final layer with sigmoid activation. This model is flexible and can capture complex patterns optimized for binary classification.

The Fig. 2(a) and Fig. 2(b) represents the general flow in which the web-based model interface for the prediction of the various diseases and shows the user interaction with the model. These figures also represent the various models implemented for the prediction of the diseases. The Fig. 2(c) additionally represents the neural network model summary implemented for the study and prediction of Parkinson's disease.

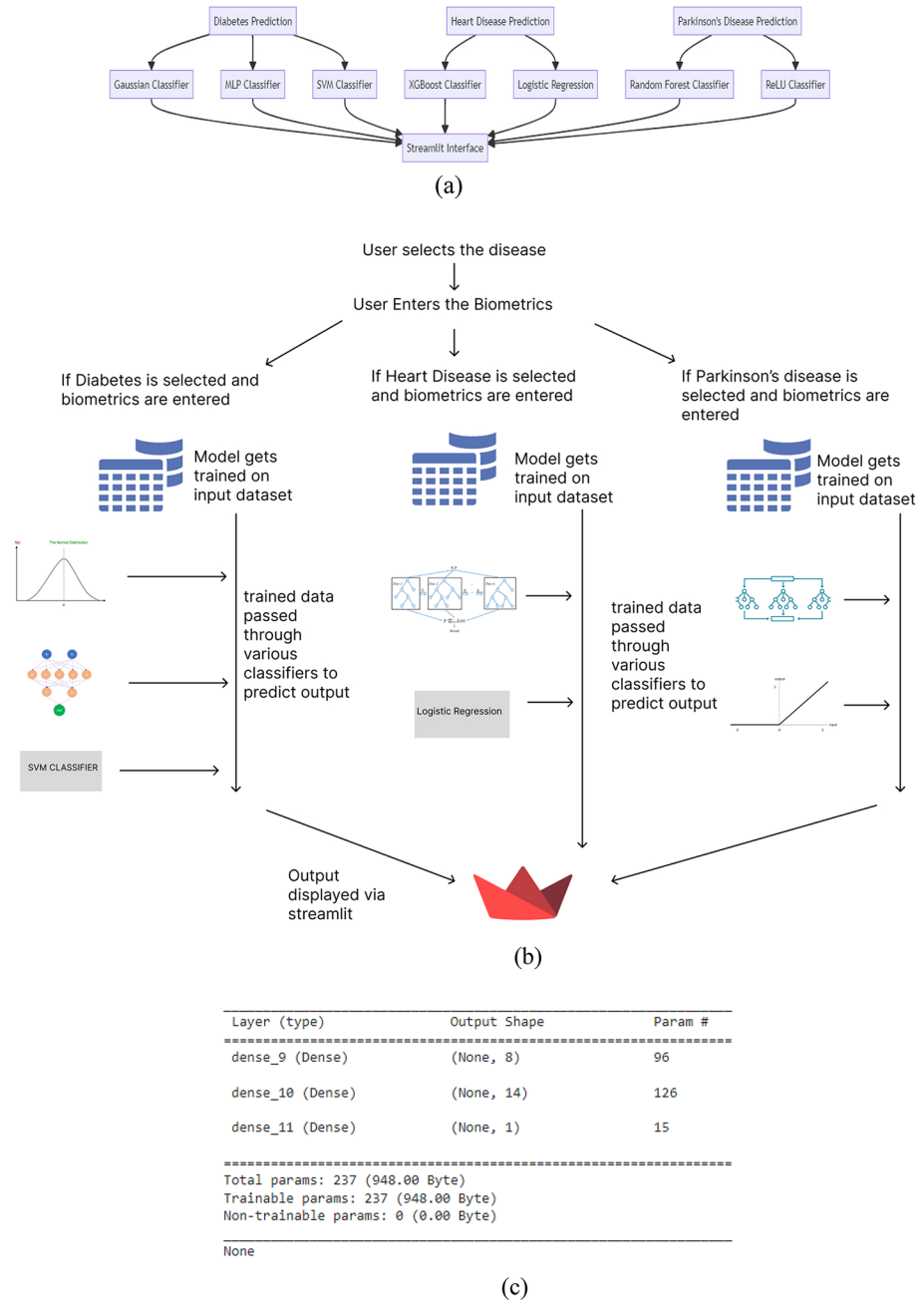


Fig. 2 (a) Flowchart for predicting the diseases using Streamlit platform (b) Detailed description for the working model (c) Parkinsons model summary after applying ReLU

XGBooster: Extreme Gradient Boosting (XGBoost), a more effective and refined implementation of the Gradient Boosting algorithm, is like the gradient boosting framework. It has also been widely used in machine learning competitions due to its strong accuracy. It is an ensemble learning technique that strengthens predictions by combining those of several weak models [17].

Random forest classifier: An ensemble machine learning model called a Random Forest Classifier is used for both classification and regression tasks. It is a development of decision trees, which are straightforward yet effective models for forecasting. Random Forests increase predictive accuracy while decreasing overfitting, which overcomes decision trees' drawbacks. [18]

Support vector machine: Support vector machines, also known as SVMs, are a very popular class of classification models. In addition to handling computationally demanding data sets, nonlinearly separable data classification, and complex machine learning problems, they are superior to other classification problems in several ways [19].

To increase the distance between the two categories, SVM translates training examples to spatial positions. New samples are then projected into that same area and predicted to fall into a category based on which side of the gap they fall on. To calculate the (soft margin) SVM classifier, an expression of the form is

$$\left[\frac{1}{n} \sum_{i=1}^n \max(0, 1 - y_i (w^T x_i - b)) \right] + \lambda ||w||^2 \quad (1)$$

If the training set can be linearly divided into two classes, one can select two parallel hyperplanes that divide the two groups, reducing the distance between them. The maximum-margin hyperplane is the one that is located halfway between these two hyperplanes. This region is referred to as the "margin" between these two hyperplanes. These hyperplanes can be defined by equations using a normalized or standardized dataset.

$$w^T x - b = 1 \text{ and } w^T x - b = -1 \quad (2)$$

Input: training set $\{(x_i, y_i)\}_i^N$, a differentiable loss function $L(y, F(x))$, several weak learners M and a learning rate α . Includes features such as glucose levels, blood pressure, BMI, age, etc., and a target variable indicating whether a patient has diabetes (1) or not (0).

Algorithm:

Initialize the model with a constant value:

$$\hat{f}_{(0)}(x) = \operatorname{argmin} \sum_{i=1}^N L(y_i, \theta) \quad (3)$$

A binary cross-entropy loss function, which measures the difference between predicted probabilities and actual outcomes.

For $m = 1$ to M :

Use the gradients and Hessians

$$\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x)=\hat{f}_{(m-1)}(x)} \quad (4)$$

$$\hat{h}_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial^2 f(x_i)} \right]_{f(x)=\hat{f}_{(m-1)}(x)} \quad (5)$$

Computing gradients and Hessians of the loss function with respect to the predictions to guide the training of each weak learner.

Fit the base learner or the weak learner using the training set

$$\left\{ x_i, -\frac{\hat{g}(x_i)}{\hat{h}(x_i)} \right\} \quad i = 1 \quad N \quad (6)$$

by solving the optimization problem below:

$$\hat{\phi} = \underset{\phi}{\operatorname{argmin}} \sum_{i=1}^N \frac{1}{2} \hat{h}(x_i) \left[-\frac{\hat{g}(x_i)}{\hat{h}(x_i)} - \phi(x_i) \right]^2 \quad (7)$$

$$\hat{f}(x) = \alpha \hat{\phi}(x) \quad (8)$$

Training each decision tree on the dataset, adjusting its parameters to minimize the loss function.

Update the Model:

$$\hat{f}_{(m)}(x) = \hat{f}_{(m-1)}(x) + \hat{f}_{(m)}(x) \quad (9)$$

Sequentially add each decision tree's predictions to the ensemble, weighted by the learning rate.

Output from the above operations is:

$$\hat{f}(x) = \hat{f}_{(M)}(x) = \sum_{m=0}^M \hat{f}_{(m)}(x) \quad (10)$$

A gradient boosting model that can predict the likelihood of a patient having diabetes based on their features.

4 Results and discussion

This section provides the results using an interactive front end where the user can enter the data in the required field and get the result whether the person is positive of the disease or negative. The front end is created using Streamlit which provide a user-friendly interface and is assisted with interactive graph which is created using the chart and graph feature of

the Streamlit and helps to visualize the changes. In the back end we load the model pre-trained with convolutional neural network using pickle and predict result dynamically by taking input from the user and feeding it to the neural network model and getting the result.

The evaluation criteria for the model's performance, which are observed by the confusion matrix and are thus presented in Table 5. The Table 6 further provides the confusion matrix metrics for the various models implemented and provides a comparative analysis for estimating the and understand the results obtained. Through the confusion matrices in Figs. 3(a-c), we can learn about the usage class of testing data in the model.

Accuracy: The model's precision of 98% is noticeable, i.e., it correctly distinguishes 98% of the instances.

Precision and Recall: Both metrics perform excellently at 90.32%, suggesting that the model can express both positive and negative predictions.

Specificity: The model specific training session is the specificity of 88.9%, certainly precise in defining all cases.

From Table 6 we get the following key observations on the predictions of various diseases considered under the scope of the paper.

- **Diabetes:** The Support Vector Machine classifier achieved the highest precision and F1-score, indicating it handles this binary classification problem more effectively than the MLP and Gaussian Process classifiers.
- **Heart Disease:** Logistic Regression showed the highest specificity and MCC, making it the most reliable for predicting heart disease in this comparison.
- **Parkinson's Disease:** The Random Forest classifier demonstrated perfect sensitivity but low precision, suggesting it correctly identifies all positive cases but has a high rate of false positives. ReLU performed well with balanced sensitivity and specificity.

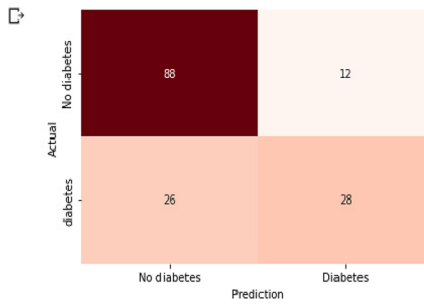
Table 5 Confusion matrix measures for diabetes, heart disease and parkinson's disease

Measures	Diabetes	Heart disease	Parkinson's disease	Derivations
	Values			
Sensitivity	0.5185	0.78	0.9032	$TPR = TP / (TP + FN)$
Specificity	0.88	0.7561	0.652	$SPC = TN / (FP + TN)$
Precision	0.7	0.7959	0.9032	$PPV = TP / (TP + FP)$
Negative Predictive Value	0.7719	0.7381	0.625	$NPV = TN / (TN + FN)$
False Positive Rate	0.12	0.2439	0.375	$FPR = FP / (FP + TN)$
False Discovery Rate	0.3	0.2041	0.09068	$FDR = FP / (FP + TP)$
False Negative Rate	0.4815	0.22	0.09068	$FNR = FN / (FN + TP)$
Accuracy	0.7532	0.7692	0.8462	$ACC = (TP + TN) / (P + N)$
F1 Score	0.5957	0.7879	0.9062	$F1 = 2TP + FP + FN$
Matthews Correlation Coefficient	0.4337	0.5351	0.5282	$TP * TN - FP * FN / \sqrt{((TP + FP) * (TP + FN) * (TN + FP) * (TN + FN))}$

Table 6 Comparison of CM metrics for various models implemented for disease prediction

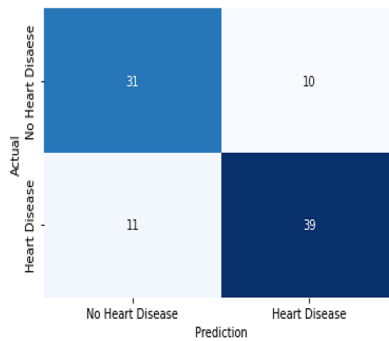
Disease & Models Parameters	Diabetes			Heart Disease		Parkinson's Disease	
	MLP Classifier	Gaussian Process Classifier	Support Vector Machine	Logistic Regression	XGB Classifier	Random Forest	ReLU
Sensitivity	0.7719	0.7542	0.7778	0.7931	0.7442	1	0.75
Specificity	0.07	0.6944	0.7568	0.8438	0.8125	0.8611	0.8571
Precision	0.88	0.89	0.91	0.8214	0.7805	0.357	0.375
Negative Predictive Value	0.5185	0.463	0.5185	0.8182	0.78	1	0.9677
False Positive rate	0.3	0.3056	0.2432	0.1563	0.1875	0.1389	0.1429
False Discovery rate	0.12	0.11	0.09	0.1786	0.2195	0.625	0.625
False Negative Rate	0.2281	0.2458	0.2222	0.2069	0.2558	0	0.25
Accuracy	0.7532	0.7403	0.7727	0.8197	0.7802	0.8718	0.8462
F1-score	0.8224	0.8165	0.8387	0.807	0.7619	0.5455	0.5
Matthew's Correlation Coefficient	0.4337	0.398	0.4786	0.6382	0.5586	0.5683	0.4562

Fig. 3 (a) Confusion matrix for Diabetes (b) Confusion matrix for Heart Disease (c) Confusion matrix for Parkinson's Disease

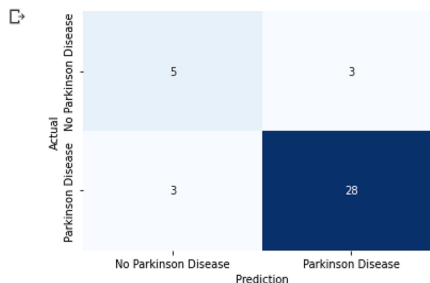


<Figure size 432x288 with 0 Axes>

(a)



(b)



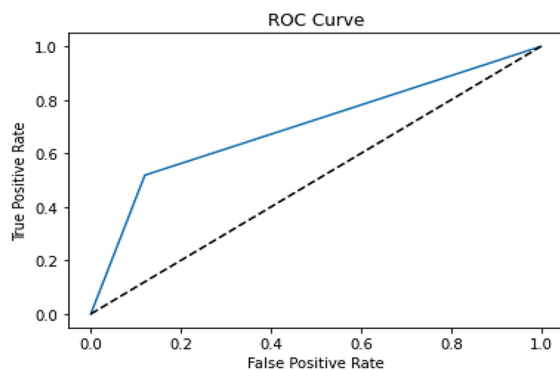
<Figure size 432x288 with 0 Axes>

(c)

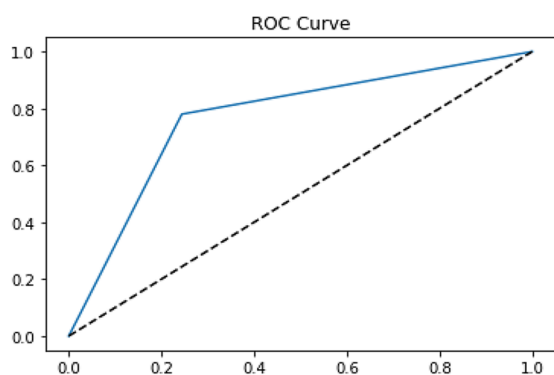
Overall, the choice of the classifier significantly impacts the performance metrics across different diseases, highlighting the importance of selecting an appropriate model based on specific requirements and characteristics of the dataset.

Furthermore, the ROC curves plotted in figures 4(a-c), defined and used to assess the overall performance of the model. It also helps us to understand the efficiency of the model deployed for the prediction of the diseases. As the below curves are close to the top-left

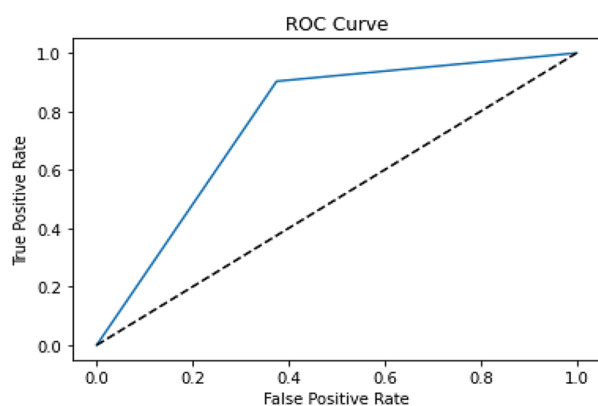
Fig. 4 (a) ROC for Diabetes (b) ROC for Heart Disease (c) ROC for Parkinson's Disease



(a)

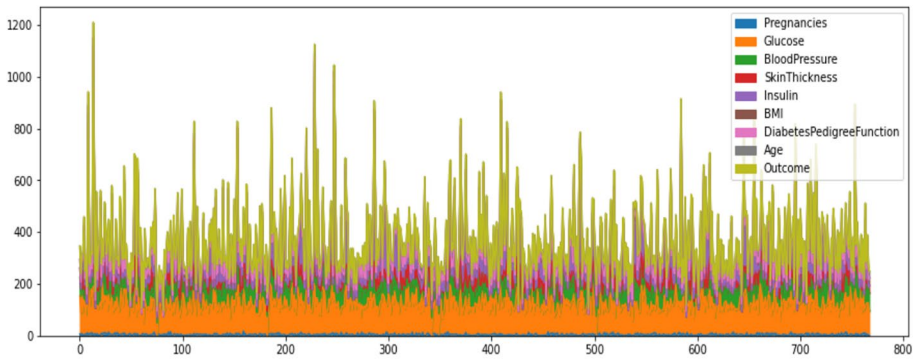


(b)



(c)

corner i.e., the ideal clinical discriminator and increases sharply during the model evaluation. One can conclude that the given models perform well on unseen or new data and that the model is neither under-fitting or over-fitting.



(a)

Diabetes Prediction using ML



(b)

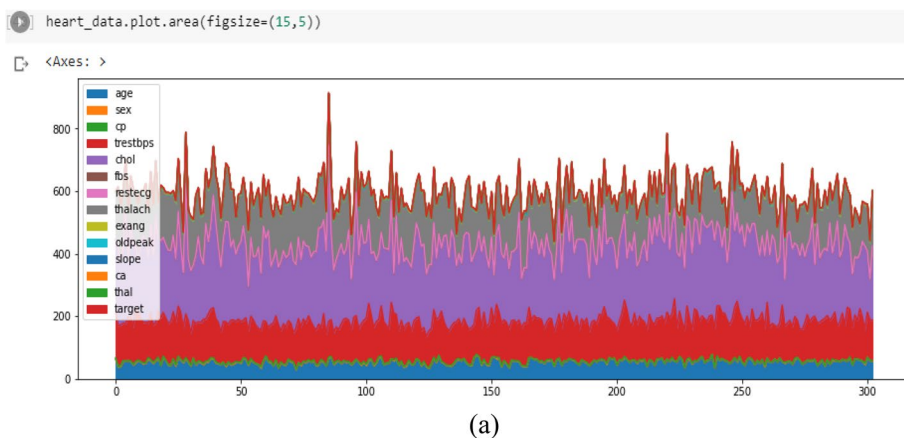
Fig. 5 (a) Data Visualization after cleaning (b) User Interface for Diabetes Prediction

Data visualization after cleaning and user interface for diabetes prediction with their biomarkers are given in figure (5a) and figure (5b), respectively. Figure (6a) represents biomarkers of heart diseases after cleaning and figure (6b) shows user interface for heart disease prediction with their biomarkers. Biomarkers of Parkinson's disease are visualized in figure (7a) after cleaning. Figure (7b) shows GUI of Parkinson's disease prediction using ML.

Result Visualization for the prediction of diabetes is given in figure (8)

The developed prediction system shows the following accuracies:

- Diabetes Prediction System: 75.32%
- Heart Disease Prediction System: 76.92%
- Parkinson's Disease Prediction System: 84.62%



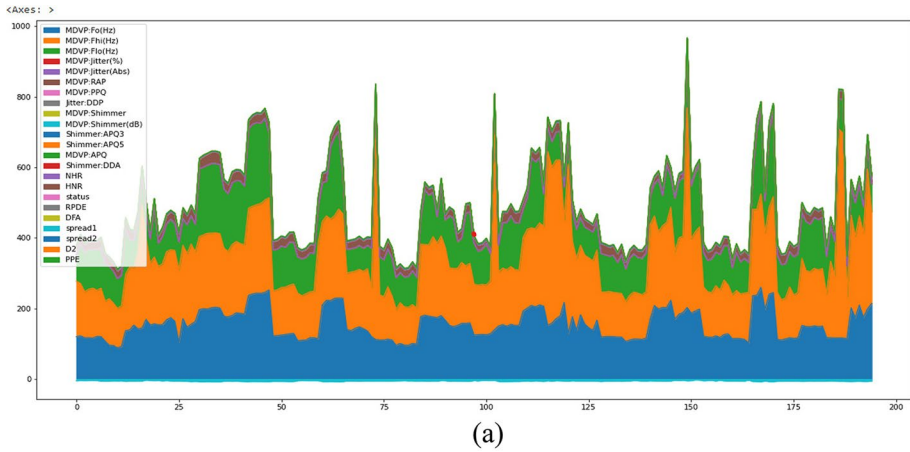
Heart Disease Prediction using ML

Age	Sex	Chest Pain types
Resting Blood Pressure	Serum Cholesterol in mg/dl	Fasting Blood Sugar > 120 mg/dl
Resting Electrocardiographic results	Maximum Heart Rate achieved	Exercise Induced Angina
ST depression induced by exercise	Slope of the peak exercise ST segment	Major vessels colored by flourosopy
thal: 0 = normal; 1 = fixed defect; 2 = reversable defect		
Heart Disease Test Result		

(b)

Fig. 6 (a) Data Visualization after cleaning (b) User Interface for Heart Disease Prediction

Overall, the results and the test set performance metrics imply that the models are training and hyperparameters tuning went better than expected. The presence of good accuracy, precision, specificity, recall for all the models signifies the ability of the models to generalize while predicting diseases based on the input of the users. These models are also



Parkinson's Disease Prediction using ML

MDVP:F0(Hz)	MDVP:F1(Hz)	MDVP:F2(Hz)	MDVP:F3(Hz)	MDVP:F4(Hz)
MDVP:F5(Hz)	MDVP:F6(Hz)	MDVP:F7(Hz)	MDVP:F8(Hz)	MDVP:F9(Hz)
MDVP:F10(Hz)	MDVP:F11(Hz)	MDVP:F12(Hz)	MDVP:F13(Hz)	MDVP:F14(Hz)
MDVP:F15(Hz)	MDVP:F16(Hz)	MDVP:F17(Hz)	MDVP:F18(Hz)	MDVP:F19(Hz)
MDVP:F20(Hz)	MDVP:F21(Hz)	MDVP:F22(Hz)	MDVP:F23(Hz)	MDVP:F24(Hz)
MDVP:F25(Hz)	MDVP:F26(Hz)	MDVP:F27(Hz)	MDVP:F28(Hz)	MDVP:F29(Hz)
MDVP:F30(Hz)	MDVP:F31(Hz)	MDVP:F32(Hz)	MDVP:F33(Hz)	MDVP:F34(Hz)
MDVP:F35(Hz)	MDVP:F36(Hz)	MDVP:F37(Hz)	MDVP:F38(Hz)	MDVP:F39(Hz)
MDVP:F40(Hz)	MDVP:F41(Hz)	MDVP:F42(Hz)	MDVP:F43(Hz)	MDVP:F44(Hz)
MDVP:F45(Hz)	MDVP:F46(Hz)	MDVP:F47(Hz)	MDVP:F48(Hz)	MDVP:F49(Hz)
MDVP:F50(Hz)	MDVP:F51(Hz)	MDVP:F52(Hz)	MDVP:F53(Hz)	MDVP:F54(Hz)
MDVP:F55(Hz)	MDVP:F56(Hz)	MDVP:F57(Hz)	MDVP:F58(Hz)	MDVP:F59(Hz)
MDVP:F60(Hz)	MDVP:F61(Hz)	MDVP:F62(Hz)	MDVP:F63(Hz)	MDVP:F64(Hz)
MDVP:F65(Hz)	MDVP:F66(Hz)	MDVP:F67(Hz)	MDVP:F68(Hz)	MDVP:F69(Hz)
MDVP:F70(Hz)	MDVP:F71(Hz)	MDVP:F72(Hz)	MDVP:F73(Hz)	MDVP:F74(Hz)
MDVP:F75(Hz)	MDVP:F76(Hz)	MDVP:F77(Hz)	MDVP:F78(Hz)	MDVP:F79(Hz)
MDVP:F80(Hz)	MDVP:F81(Hz)	MDVP:F82(Hz)	MDVP:F83(Hz)	MDVP:F84(Hz)
MDVP:F85(Hz)	MDVP:F86(Hz)	MDVP:F87(Hz)	MDVP:F88(Hz)	MDVP:F89(Hz)
MDVP:F90(Hz)	MDVP:F91(Hz)	MDVP:F92(Hz)	MDVP:F93(Hz)	MDVP:F94(Hz)
MDVP:F95(Hz)	MDVP:F96(Hz)	MDVP:F97(Hz)	MDVP:F98(Hz)	MDVP:F99(Hz)
Parkinson's Test Result				

(b)

Fig.7 (a) Data Visualization after cleaning (b) User Interface for Parkinson's Disease Prediction

robust to outliers and incorrect data, that means they will make sound predictions of diseases. In general, a well-designed backend model complemented with the interactive front end acts an effective tool to predict diseases such as Diabetes, Heart Disease, Parkinson disease which provides a rich user experience due to a more understandable and realistic prediction.

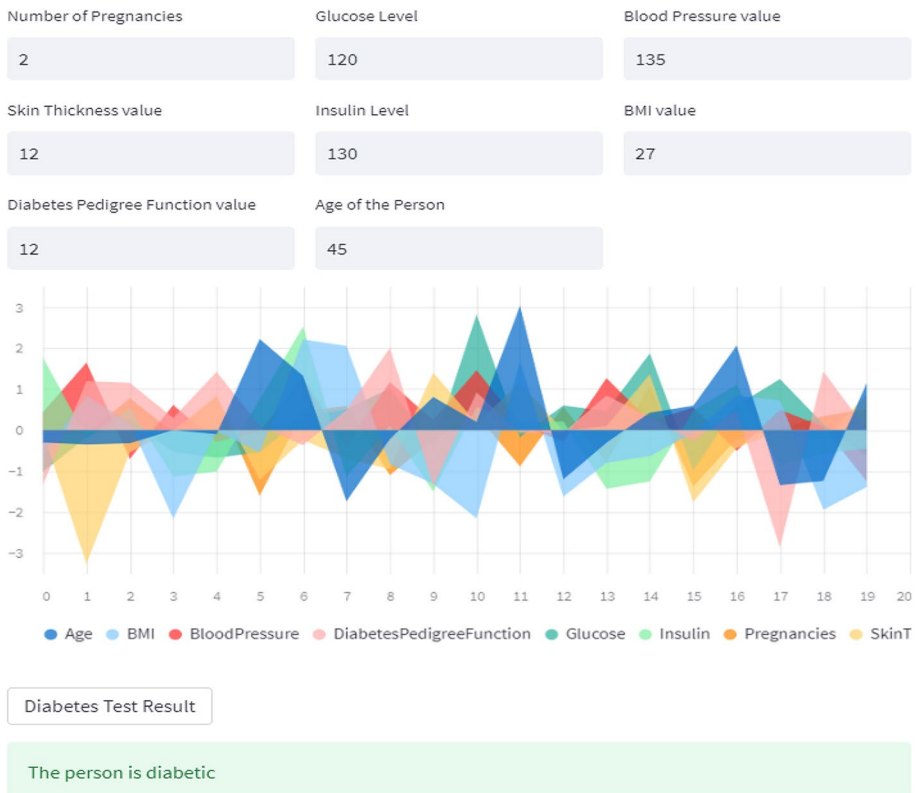


Fig. 8 Result Visualization for the prediction

5 Conclusion and future scope

With the fast-moving world and everyday innovation in a newer technology for easing the human life, this paper proposes a prediction system of lifestyle diseases using the machine learning algorithms. Along with increase stress levels of people and poor mental health, have made such early disease predicting software a need of the time. Prediction system has been developed keeping in mind the need of the time i.e., to build a healthy and disease-free society and its accurate prediction system will, in future, lead to contribution in the field of medical science for early prediction of lifestyle and sedentary diseases and help in early treatment thereby in return saving lives and making this society a better place to live in. This will also help in achieving one of the United Nations' 17 objectives is to ensure healthy lives and promote the well-being of all people at all ages.

Funding No funding is available for the submitted manuscript.

Data availability The datasets generated or analyzed during the current study are available from.
<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>
<https://www.kaggle.com/datasets/thecansin/parkinsons-data-set>
<https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci>

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

- Kulzhanova ZT, Kulzhanova GT, Mukhanbetkaliyev YY, Abdildina MKKKS (2020) Impact of technology on modern society—a philosophical analysis of the formation of technogenic environment. *Media Watch* 11(3):537–549
- Kulzhanova ZT, Kulzhanova GT, Mukhanbetkaliyev YY, Kakimzhanova MK, Abdildina KS (2020) Impact of technology on modern society—a philosophical analysis of the formation of technogenic environment. *Media Watch* 11(3):537–549
- Mathur P, Mascarenhas L (2019) Lifestyle diseases: Keeping fit for a better tomorrow. *Indian J Med Res* 149(Suppl 1):S129
- Roth G (2018) Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2017 (GBD 2017) Results. Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2018. *The Lancet* 392:1736–1788
- “What’s behind the rise in heart attacks among young people?,” *Cminj.com*. [Online]. Available: <https://www.cminj.com/blog/whats-behind-the-rise-in-heart-attacks-among-young-people>. [Accessed: 17-Jun-2023].
- “Parkinson’s disease: Causes, symptoms, and treatments,” National Institute on Aging. [Online]. Available: <https://www.nia.nih.gov/health/parkinsons-disease>. [Accessed: 17-Jun-2023].
- Alić B, Gurbeta L, & Badnjević A (2017) Machine learning techniques for classification of diabetes and cardiovascular diseases. *IEEE*. 1–4.
- Yaganteeswarudu A. (2020) Multi disease prediction model by using machine learning and Flask API. *IEEE*. 1242–1246.
- Krishnamoorthi R, Joshi S, Almarzouki HZ, Shukla PK, Rizwan A, Kalpana C, & Tiwari B (2022) *Journal of Healthcare Engineering*, 2022.
- Dhanka S, & Maini S (2021) Random Forest for Heart Disease Detection: A Classification Approach. *IEEE*. 1–3.
- Abdellatif A, Abdellatif H, Kanesan J, Chow CO, Chuah JH, Gheni H (2022) An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyperparameter Optimization Methods. *IEEE Access* 10:1–1. <https://doi.org/10.1109/ACCESS.2022.3191669>
- Chaki J, Ganesh ST, Cidham SK, Theertan SA (2022) Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. *J King Saud Univ Comput Inform Sci* 34(6):3204–3225. <https://doi.org/10.1016/j.jksuci.2020.06.013>
- Arumugam K, Naved M, Shinde PP, Leiva-Chauca O, Huaman-Osorio A, Gonzales-Yanac T (2023) Multiple disease prediction using Machine learning algorithms. *Mater Today: Proceedings* 80:3682–3685
- Gupta J, Pathak S, Kumar G (2022) A hybrid optimization-tuned deep convolutional neural network for bare skinned image classification in websites. *Multimed Tools Appl* 81(18):26283–26305
- Gupta J, Pathak S, Kumar G (2022) Aquila coyote-tuned deep convolutional neural network for the classification of bare skinned images in websites. *Int J Mach Learn Cybern* 13:3239–3254
- Abdalrada AS, Abawajy J, Al-Quraishi T, Islam SMS (2022) Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: a retrospective cohort study. *J Diabetes Metab Disord* 21(1):251–261
- Wikipedia contributors, “XGBoost,” Wikipedia, The Free Encyclopedia, [Online]. Available: <https://en.wikipedia.org/w/index.php?title=XGBoost&oldid=1160156324>. [Accessed]14-Jun-2023.
- Wikipedia contributors, “Random forest,” Wikipedia, The Free Encyclopedia,[Online]. Available: https://en.wikipedia.org/w/index.php?title=Random_forest&oldid=1160091469. [Accessed]14-Jun-2023.
- Wikipedia contributors, “Support vector machine,” Wikipedia, The Free Encyclopedia, [Online]. Available: https://en.wikipedia.org/w/index.php?title=Support_vector_machine&oldid=1156332283[Accessed]22-May-2023

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.