Explore–Journal of Research
Peer Reviewed Journal
ISSN 2278 – 0297 **(Print)**
ISSN 2278 – 6414 **(Online)**
© Patna Women's College, Patna, India
https://www.patnawomenscollege.in/journal

HEALTHMATRIX: Multiple Disease Prediction System

• **Manisha Prasad** • **Aanshi Kumari** • **Aradhana Prakash**
• **Kumari Shreya Arvind**

Received : October, 2025
Accepted : January, 2026
Corresponding Author : Manisha Prasad

Abstract: *As per World Health Organization mostly deaths were caused by chronic diseases worldwide in the year 2021. Bad lifestyle, late diagnosis, and limited diagnosis tools leads to increasing global deaths and lifetime health issues.So, to prevent worse health conditions early prediction of diseases is necessary.*

In order to determine the risk of four major non-communicable diseases diabetes, liver disease, heart disease, and brain stroke, this study suggests a machine learning-based multiple illness prediction system. The study uses public available health data. Graphs and Algorithms are used to test the prediction performance. To ensure authenticity of prediction system, performance of algorithms was tested using different metrics.

The method showed how prediction systems with machine learning application can help doctors make early identification of disease condition, encourage early treatment, and decrease the increasing impact of chronic diseases worldwide.

Keywords: *Algorithms; Diabetes; Early Diagnosis; NCD, Predictive Models; Medical Data Analysis.*

Manisha Prasad
Head, Department of Computer Applications (BCA)
Patna Women's College (Autonomous),
Bailey Road, Patna–800 001, Bihar, India
E-mail: manisha.bca@patnawomenscollege.in

Aanshi Kumari
BCA III year, Session: 2023-2027,
Patna Women's College (Autonomous),
Patna University, Patna, Bihar, India

Aradhana Prakash
BCA III year, Session: 2023-2027,
Patna Women's College (Autonomous),
Patna University, Patna, Bihar, India

Kumari Shreya Arvind
BCA III year, Session: 2023-2027,
Patna Women's College (Autonomous),
Patna University, Patna, Bihar, India

Introduction:

Due to availability of healthcare data, new ways are created to use ML in identifying medical conditions. ML algorithms are able to understand medical data more accurately and in a faster way than traditional methods, which can help doctors in identifying diseases at an early stage. Most of the deaths worldwide are caused by these major non-communicable diseases. There are still some regions which lack medical facilities that can lead to late detection of disease. Basically, this system is defined for both health care professionals and individuals for predicting the diseases with accuracy so that further decision could be taken by them.

Traditional diagnostic processes can be slow, expensive, and highly dependent on expert interpretation, and they often struggle to analyse medical information that include multiple health-related features. In contrast, ML-based systems can smoothly handle more number of data, find hidden patterns, and

detect even minor risk indicators that may not be visible through any person's manual evaluation. Motivated by these challenges, this research develops a **Multiple Disease Prediction System** that predicts the risk of four major chronic diseases using separate ML algorithms for each disease. Algorithms are trained on publicly available datasets to achieve accurate performance.

Different health features have been used for analysis of various medical factors which helps in seeing different patterns and identifying the risk.Thus, this system supports early health awareness depending on data that has been processed.

Objectives:

- To develop an automated machine learning based system that can identify multiple diseases such as Diabetes, Heart Disease, Liver Disease, and Stroke using clinical input data for early identification and diagnosis.
- To advance existing prediction approaches by providing a single platform that supports multiple disease predictions individually, instead of relying on separate and isolated systems for each disease.
- To help patients and doctors by providing accurate and easy to understand prediction results. To allow them to make early medical decisions. So, that chances of serious conditions can be reduced.
- To design a user-friendly graphical interface (GUI) in which users can enter medical details. So, that users can receive instant results of prediction and makes the process of prediction easy for healthcare professionals and individuals.

The objective of such an intelligent system is to increase human intelligence by providing accurate detection of disease. It can help individuals and healthcare professionals in making decisions in time.

Hypothesis:

1. Diabetes Prediction

Problem Identified: The prediction of diabetes with the help of early diabetes symptoms can be difficult for individuals. This causes late diagnosis and serious conditions.

Proposed Solution: To develop a prediction system that is based on machine learning and to know the early signs of diabetes through patient health details.

2. Heart Disease Prediction

Problem Identified: The prediction of heart disease with the help of traditional methods can be complex and can lead to late in treatment. Early symptoms of heart disease are often missed by patients.

Proposed Solution: To use Machine Learning(ML) algorithms to check health parameter such as cholesterol, heart rate and other parameters.

3. Liver Disease Prediction

Problem Identified: In the early stage, liver disease has some or no symptoms. Patients and Doctors face difficulty to detect and control before serious condition.

Proposed Solution: To use Machine Learning based prediction model to check health parameters such as bilirubin, enzyme levels, albumin, and protein values. This can help doctors to detect early chances of liver disease.

4. Brain Stroke Prediction

Problem Identified: Stroke is unpredictable and is identified only when major symptoms are checked, resulting in high mortality and disability rates.

Proposed Solution: To use Machine Learning based prediction model to check health parameters such as hypertension, smoking habits, obesity, glucose level, and age to know stroke risk and help people get early medical treatment.

Proposed Model:

The aim of this research is to help early prediction of multiple chronic diseases using Machine Learning (ML) techniques. The given system is designed to check the risk of diabetes, heart disease, liver disease, and brain stroke using patient health data.

The Multiple Disease Prediction System uses trained machine learning models to study patient parameters and give prediction outcomes. Different

models are developed for each disease. So, that the model becomes more accurate and trustworthy.

The system provides simple graphic interface in which user can enter medical data and get prediction results as Positive or Negative. This helps users to detect early risks. The system tells how machine learning can help in early disease prediction.

Barriers to Identifying Diseases at an Early Stage: Non communicable diseases often get unnoticed until the conditions become serious. Doctors are overloaded with work and have limited time, which makes it difficult for them to analyze complex medical data on their own. By identifying hidden patterns in large dataset ML helps in detecting the possibility of a disease early. It also provides quick, data related insights which help professionals in their decisions. When combined with human expertise, these systems can play major role in improving detection, guide better treatments and reduce the deaths caused by these chronic diseases.

Machine-Learning (ML): Machine learning comes under artificial intelligence i.e it is a part of AI that helps systems learn from the data provided to them without giving step by step rules. ML plays an important role in healthcare as it provides doctors in making decisions quickly and identifying risks early. Thus reducing manual human-made mistakes. The two ways in which pattern detection is done : Firstly, they learn from data that has labels . Secondly, from data that has no labels. After training and testing the systems become much more trustworthy. Thus, ML is widely used for real-world predictions.

AI is the broader field of intelligent systems, while ML focuses on data-based learning, and DL uses human-inspired learning network to learn complex patterns.

1. Diabetes Prediction Model: PIMA Indians Diabetes Dataset comprising 768 patient records was used for this type of disease prediction(Smith *et al.*,1988).SMOTE was used to remove biasness in the data(Chawla *et al.*, 2002). Feature engineering included BMI Category, Glucose Category, and Insulin Score derived from established medical guidelines, with One-Hot Encoding and Robust Scaling applied for feature normalization (Pedregosa *et al.*, 2011). Comparative evaluation revealed that the Random Forest achieved the maximum F1-score of 0.84, which shows a balanced trade-off between recall and precision, and was therefore selected as the final prediction model (Breiman, 2001).

Table 1. Diabetes Prediction model

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest (RF)	0.8766	0.81	0.91	0.84
Logistic Regression (LR)	0.8831	0.75	0.91	0.83
Decision Tree	0.8636	0.72	0.89	0.80
K-Nearest Neighbors (KNN)	0.8441	0.71	0.83	0.76

From the comparison in the table above, the first algorithm achieved the highest F1-score, Since F1-score is the most reliable metric for imbalanced data, this result confirms that the algorithm is most suitable for the prediction.

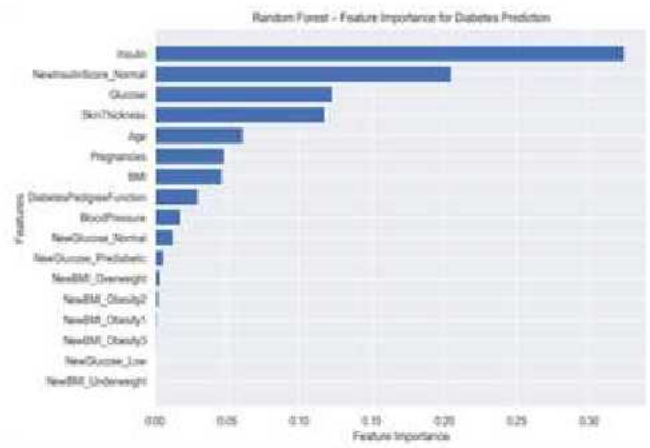


Fig. 1. Was used to identify the most helpful predictors in the Random Forest algorithm based on Gini impurity reduction. Insulin emerged as the dominant feature, contributing over 30% to the model's decisions, followed by the engineered features New Insulin Score, New Glucose, and New BMI. Other factors such as BMI, Age, Pregnancies, and Diabetes Pedigree Function also showed meaningful influence. These results confirm that these features make the model's predictions

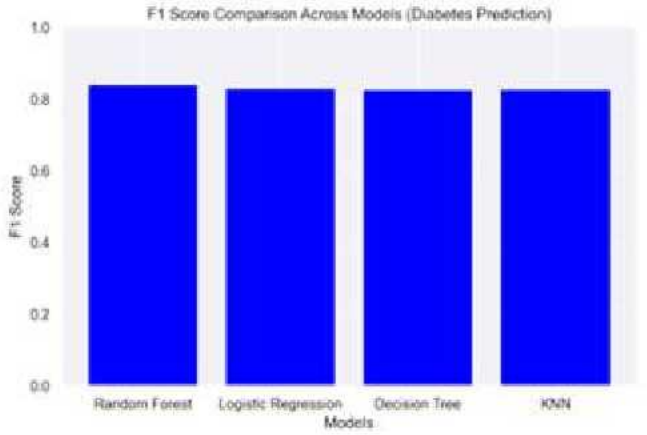


Fig. 2. The evaluation showed only small differences in performance among all four models, with nearly similar F1-scores. However, Random Forest was selected as the final model because it gives more consistent and reliable predictions on new, unseen data

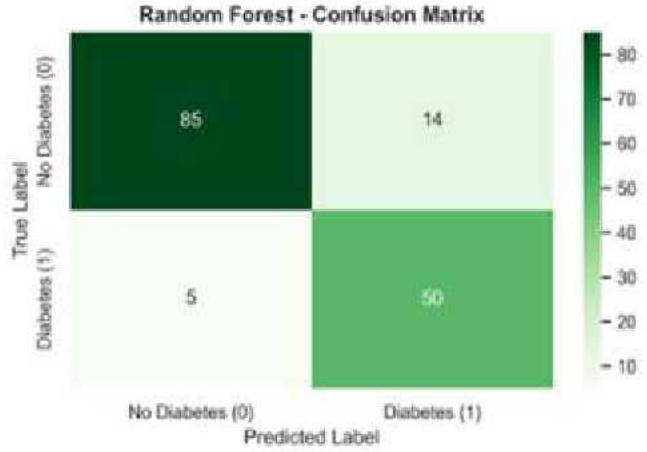


Fig. 3: Shows that the algorithm accurately identified 85 without disease cases (TN) and 50 with disease cases (TP), while 14 non-diabetic cases were misclassified as diabetic (FP) and 5 diabetic cases were missed (FN). The relatively low number of (FN = 5) demonstrates the algorithm's importance in minimizing missed cases

2. Model for Liver Disease Detection: The Indian Liver Patient Dataset (ILPD) dataset was used for the early prediction of this disease (Frank,2010). Pre-processing of data involved label encoding, Standard Scaler for feature scaling, and SMOTE for class imbalance correction(Chawla *et al.,* 2002).Different algorithm was used to evaluate the model. (Pedregosa *et al.*, 2011).Due to the dataset's imbalance and nonlinear patterns, performance was assessed using the F1 score (Pedregosa *et al.*, 2011).XGBoost achieved the best balance between precision and recall and was selected as the final model (Chen, 2016).

Table 2. Liver Disease Prediction

Algorithm	Accuracy	Precision	Recall	F1 Score
XG Boosting	0.844311	0.870130	0.807229	0.837500
Gradient Boosting	0.826347	0.855263	0.783133	0.817610
Random Forest	0.820359	0.863014	0.759036	0.807692
Decision Tree	0.760479	0.802817	0.680747	0.740200
Logistic Regression	0.760479	0.864407	0.614458	0.713810
KNN	0.754491	0.903846	0.566225	0.696296

As the table above illustrates, a comparative study of six machine learning algorithms was conducted. XGBoost was chosen as the final model because it had the greatest F1-Score (0.83) and the best performance in classifying liver diseases.

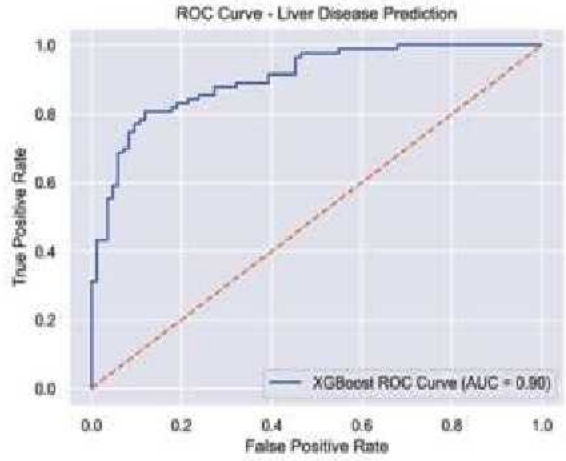


Fig. 4. This graph shows how well the liver disease algorithm works by comparing correct and wrong predictions at different levels. The graph shows clear separation between different cases, which means the algorithm works very well. With a score of 0.90, the prediction method can correctly identify unhealthy and healthy cases with high accuracy.

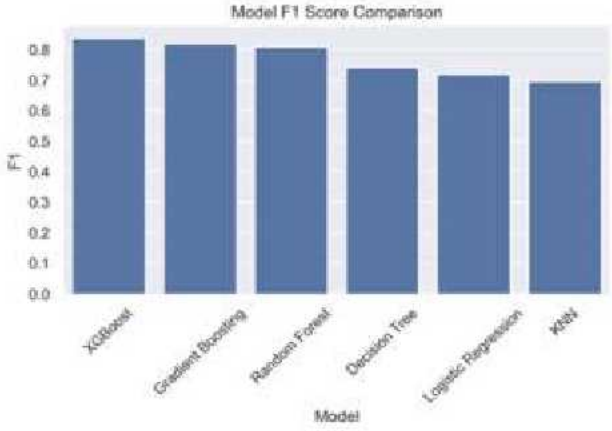


Fig. 5. The F1-Score comparison graph shows how different models performed for liver disease prediction. XGBoost has the highest F1-Score, meaning it gives a good balance between correct results and missed cases.

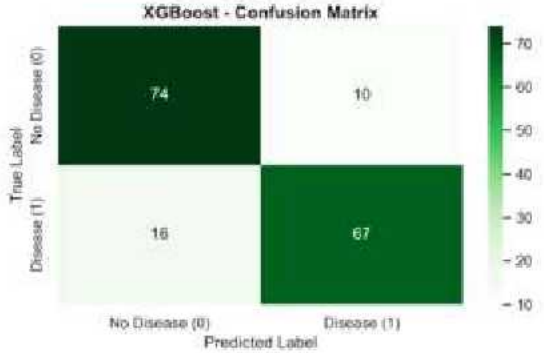


Fig. 6. Shows that the XGBoost algorithm accurately identified 74 without disease cases (TN) and 67 with disease cases (TP), with 10 (FP) and 16 (FN). The relatively low FN count indicates that the algorithm is accurate in predicting the disease cases, which is critical for timely clinical intervention.

3. Heart Disease Prediction Model: In this prediction model a supervised machine learning framework for early heart disease detection using clinical data from 1025 patients, where each patient is described by 14 medical attributes in a binary classification task (Frank, 2010). Preprocessing involved Standard Scaler-based feature normalization and an 80:20 stratified train–test split .Different metrics were used to assess five supervised classification models (Pedregosa *et al.*, 2011). Among these, the Support Vector Classifier (SVC) achieved the highest F1-score, demonstrating effective detection capability with reduced false predictions, and was therefore selected as the final algorithm (Cortes *et al.*, 1995).

Table 3: Heart Disease Prediction

Algorithm	Recall	Precision	Accuracy	F1-Score
Support Vector Classifier (SVC)	0.9429	0.9167	0.9268	0.9296
Random Forest	0.9524	0.8850	0.9122	0.9174
Decision Tree	0.8952	0.8624	0.8732	0.8785
K-Nearest Neighbors (KNN)	0.8571	0.8738	0.8634	0.8653
Logistic Regression	0.9143	0.7619	0.8098	0.8312

As shown in the comparison table, theSVC model achieved the highest F1-Score (0.93), indicating best performance in heart disease classification. Hence, SVC was selected as the final predictive model.

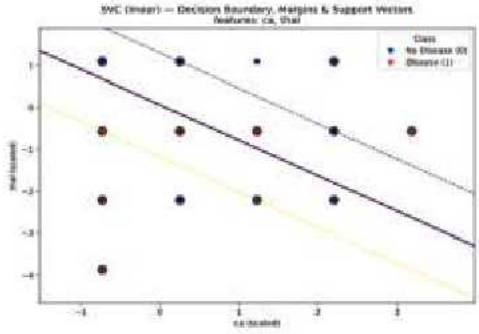


Fig. 7. illustrates the working of the SVC model using two key heart disease features, *ca* and *thal*. Blue and red points represent healthy and diseased patients, respectively. The solid line denotes the decision boundary, while the dotted lines indicate the margins.The closest points are the support vectors. The clear separation between classes highlights the discriminative power of these features and supports the strong performance of SVC in heart disease prediction.

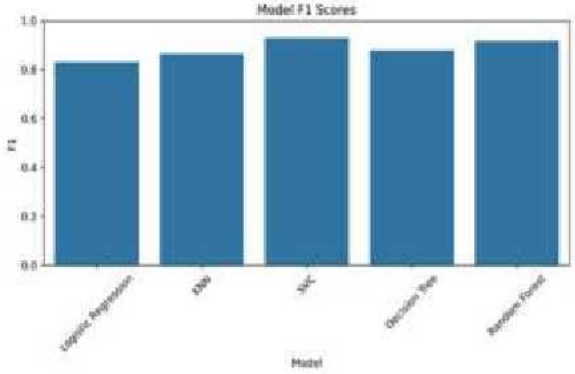


Fig. 8. The F1-Score comparison graph shows that the SVC model achieved the highest score, indicating the best balance between correctly predicting heart disease cases and avoiding misclassification.

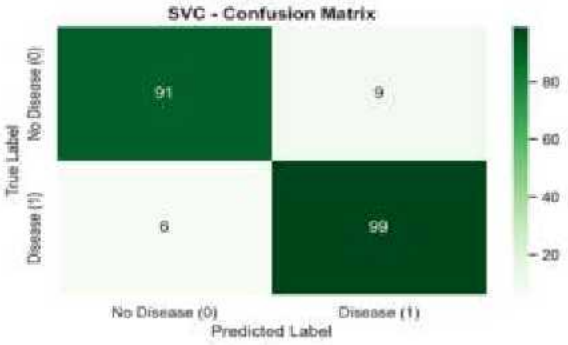


Fig. 9. shows that the SVC algorithm correctly classified 91 patients(TN) without disease and 99 patients (TP) with disease. It produced 9 (FP) and only 6 (FN). Low number of (FN) shows that the algorithm is highly accurate at predicting heart disease , which is crucial for reducing missed diagnoses in clinical practice.

4. Brain Stroke Prediction Model: This study presents a supervised machine learning method for early detection of brain stroke using 15,304 patient records comprising clinical and demographic features. SMOTE was applied to remove biasness in the data, manage missing values, encode categorical variables, and perform feature scaling (Chawla *et al.*, 2002). Different metrics were used to assess a number of classifiers.

Random Forest achieved the highest F1-score and was selected as the final prediction model, demonstrating reliable performance for early stroke risk assessment and effective clinical decision support (Breiman, 2001) .

Table 4. Brain Stroke Prediction

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9559	0.9391	0.9733	0.9559
Decision Tree	0.9346	0.9209	0.9483	0.9344
K-Nearest Neighbors (KNN)	0.9128	0.8553	0.9899	0.9177
Gradient Boosting	0.8761	0.8563	0.8988	0.8770
Logistic Regression	0.8291	0.8168	0.8408	0.8286
SVC	0.8075	0.7703	0.8665	0.8156

In the results table above, Random Forest achieved the highest performance (Accuracy = 0.9559, F1-Score = 0.9559), demonstrating strong stroke classification capability. Therefore, Random Forest was selected as the final prediction model.

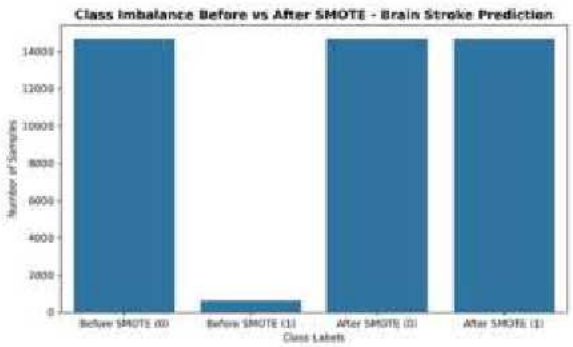


Fig. 10. The class imbalance analysis shows that before applying SMOTE, the dataset was dominated by *No Stroke* cases, with stroke instances making up less than 5%, which could bias the models. After SMOTE, the classes were balanced, allowing models to learn from both outcomes more effectively. This improved performance and helped reduce false negatives, which is critical in clinical decision.

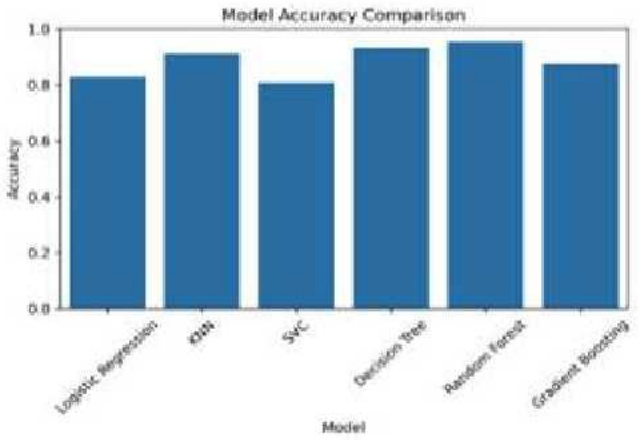


Fig. 11. Among all evaluated classifiers, Random Forest achieved the highest F1-score, effectively distinguishing between stroke and non-stroke cases, and was therefore selected as the final model for brain stroke prediction.

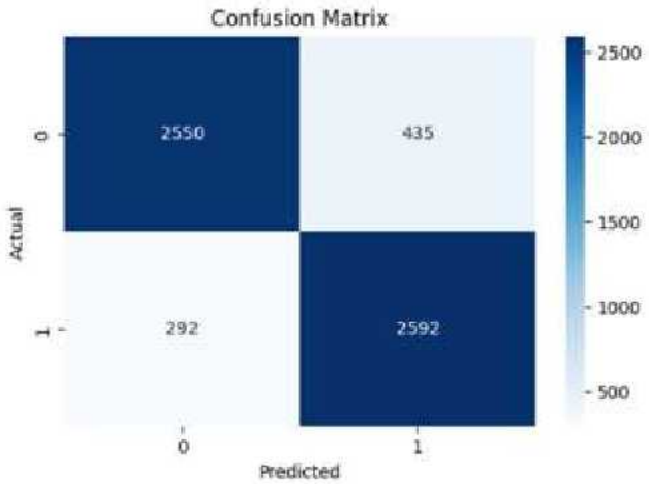


Fig. 12. The figure indicates that the Random Forest model accurately identified most stroke and non-stroke patients, with only a small number of missed stroke cases (292). This highlights the model's usefulness in clinical settings, where early and accurate stroke detection is essential.

Table 5. Best Performing Models

Disease	Best Algorithm	Reason for Selection	Performance Advantage
Diabetes Prediction	Random Forest	Achieved highest F1-Score; strong Recall for diabetic class	Handles nonlinear patterns & reduces bias using class weight.
Heart Disease Prediction	Support Vector Classifier (SVC)	Superior boundary separation & generalization	Effective in high-dimensional clinical dataset
Liver Disease Prediction	XGBoost	Best F1-Score & efficient handling of imbalance	Gradient boosting improves robustness & stability
Brain Stroke Prediction	Random Forest	Minimum error variance & stable accuracy	Ensemble averaging reduces overfitting

Future Scope:

- **Real-Time Health Data:** The program can be connected to hospital records and medical devices that are connected to the internet so that the health information of patients is updated all the time.
- **Wearable Device Integration:** Patients can use integrated wearables that check vitals all the time and send warnings if any risk is detected. These system can play important role in detecting risks early.
- **Use of Advanced Models:** CNN and LSTM can be used in future to understand complex health patterns. This will help in making the system better for understanding heart rate signals.
- **Single model for Multiple Diseases:** Having many different system we can build one system that can predict many diseases at the same time making it faster and easier for use.
- **Mobile and Cloud Access:** The system can be made available on phones or on the internet so that it is easily accessible to everyone.

Conclusion:

The study presents system that can predict these four diseases: Diabetes, Heart Disease, Liver Disease or Brain Stroke all at once. The system uses ML in making these predictions. Publically available data was used for training . It was made sure that the data on which the system is being trained is processed properly and is enough for the system to learn from. The algorithms were also tested to see how well they worked on new unseen data by using metrics like confusion matrix, ROC curves etc. For each disease the algorithm that worked the best was chosen. For example SVC performed the best and was chosen for predicting Heart Disease. Random Forest performed the best for predicting Brain Stroke. At last, all the algorithms were combined and an easy to use tkinter based interface was developed where patients can enter their health information and get prediction. In the future the system can be made better by integrating it with devices that can send warnings when risk is detected in real time using more medical information and making it available on the internet so that it is easily accessible to everyone.

References:

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., &Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Chen, T. (2016). XGBoost: A Scalable Tree Boosting System. Cornell University.

Cortes, C., &Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Frank, A. (2010). Uci machine learning repository. irvine, ca: University of california, school of information and computer science. http://archive. ics. uci. edu/ml.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... &Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988, November). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the annual symposium on computer application in medical care (p. 261).

World Health Organization. (2023). Assessing national capacity for the prevention and control of noncommunicable diseases: report of the2021 global survey. World Health Organization.