iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.59642

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

Multiple Disease Prediction System Using ML

Prof. P. Dhanwate[1], Vinit Joshi[2], Akshay Darade[3], Suyog Chaudhari[4], Surekha Bhoi[5]
[1]*Assistant Professor, Department of Computer Engineering, Sanjivani College of Engineering, Kopargaon*
[2, 3, 4, 5]*Students, Department of Computer Engineering, Sanjivani College of Engineering, Kopargaon, Ahmednagar (423601)*

Abstract: ***The increasing prevalence of diverse diseases presents a challenge to global healthcare systems, underscoring the need for innovative and efficient methods for early detection and preventive measures. This paper explores the application of machine learning algorithms in multiple disease prediction to enhance diagnostic accuracy and enable timely intervention. Leveraging diverse health-related data sources, including medical records and genomic information, comprehensive predictive models are developed. A multi-faceted machine learning approach integrates support vector machines, decision trees, neural networks, and ensemble learning methods to analyze complex data patterns. Feature selection and dimensionality reduction techniques optimize model performance and interpretability. The development of a predictive system involves essential steps such as data collection, preprocessing, and model training, followed by evaluation using metrics like accuracy and recall. Integration of Flask for web application development facilitates user interaction and prediction functionality. Deployment, testing, debugging, and ongoing maintenance ensure system efficiency and compliance with regulatory requirements for healthcare data security and privacy.***

Keywords: Multiple Disease Prediction, SVM, Diabetes Prediction, Parkinsons, Heart Disease, Kidney Disease Prediction, Feature Selection, Decision Trees, Random Forest, Flask, Model Training.

I. INTRODUCTION

In recent years, the field of machine learning has experienced remarkable progress, extending its applications across various domains, including healthcare. Amidst the rising health challenges, the demand for innovative and effective methods for disease prediction has become increasingly urgent. The ability to predict multiple diseases simultaneously using machine learning models has shown promising results in improving patient outcomes. Leveraging this capability, the development of Disease Predictor, where consumers can determine diseases based on given symptoms, marks a significant step forward. The convergence of machine learning and healthcare has paved the way for transformative predictive modeling approaches. Drawing from diverse health-related datasets, this paper delves into the realm of Multiple Disease Prediction using Machine Learning, presenting an integrative approach to revolutionize diagnostic capabilities. Unlike traditional diagnostic methods that often focus on individual diseases, our proposed framework transcends these limitations by simultaneously considering multiple health conditions. This predictive system encompasses crucial parameters for disease prediction, ensuring enhanced accuracy.

It is the major challenge in the medical sector for giving more accuracy to the peoples which will be more useful. In this project, we are going to predict diseases like Heart Disease, Parkinson's Disease, Kidney Disease and Diabetes . For the implementation of disease prediction we are going to use Machine Learning Algorithm such as SVM, Logistic Regression, Random Forest, Decision Tree, KNN, Gradient Boosting, KFold for validation, Flask for web interface. By using all these algorithms we are going to compare the accuracy of all the algorithms.

II. LITERATURE REVIEW

Literature Survey on Multiple Disease Prediction System Using Machine Learning

Author Name and Paper Name	Year	Description
Priyanka Sonar, Prof. K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches" (2019)	2019	The potential of Machine Learning methods to develop systems capable of predicting diabetes risk levels accurately, citing the utilization of Decision Tree, ANN, Naive Bayes, and SVM algorithms for model development and showcasing promising outcomes in terms of accuracy.

Archana Singh, Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms	2020	It evaluates algorithms like Support Vector Machine, Logistic Regression, Naïve Bayes, and Decision Tree, finding SVM to demonstrate the highest prediction accuracy among them. KNN gives more accuracy than other algorithms.
Harshit Gupta, Lakshay Dahiya Assistant Prof. (Dr.) Jyoti Kaushik (CSE Department)" Multiple Disease Prediction Using Machine Learning Algorithms	2022	Machine Learning techniques such as random forest, SVM, and logistic regression to develop an effective Multiple Disease Prediction System, aiming to enhance healthcare outcomes through proactive risk assessment and intervention.
Rahul R. Zaveri1, Prof. Pramila M. Chawan, "Prediction of Parkinson's Disease using Data Mining: A Survey	2020	Data mining techniques like KNN, Logistic Regression, Decision Tree, SVM, and Naive Bayes are employed to predict whether individuals exhibit symptoms of Parkinson's disease, facilitating early detection and intervention for better disease management.
S. Revathy, B. Bharathi, P. Jeyanthi, M. Ramesh, "Chronic Kidney Disease Prediction using Machine Learning Models	2019	This paper employs data preprocessing, transformation techniques, and various classifiers such as Decision Tree, Random Forest, and Support Vector Machines to develop a prediction framework for early detection of CKD, showing promising results in improving patient outcomes.

Collectively, these studies underscore the potential of machine learning in disease prediction and highlight the significance of data preprocessing, feature engineering, and model optimization in improving the accuracy and reliability of predictive models. Further research in this field is crucial to enhance the effectiveness of machine learning approaches in diagnosing and predicting various disease

III. PROPOSED SYSTEM

This section describes data generation, modelling, planning, and disease prediction. The first step is to gather information. Our planning process collected structured and unstructured data from a variety of sources. After the data is collected, it is processed and divided into maintenance data and test data. Then, the training data is trained for different periods of time using machine learning algorithms such as SVM, Decision Tree, Random Forest, Logistic Regression and KNN to increase the accuracy of the prediction. After a long time, when the desired goal is achieved, the design becomes ready for testing. In this step, the model is tested using test data to evaluate the performance of the model using ne w data that was not used for training. If the model meets the accuracy requirements of the profile test, the model is ready for export.

The proposed system that utilizes advanced machine learning techniques to predict multiple diseases simultaneously, addressing shortcomings found in current models. By integrating a variety of health data sources and employing algorithms like support vector machines, decision trees, random forest, KNN, and gradient boosting. The system ensures adaptability and can identify common risk factors shared across different health conditions. The data is first skimmed and feature selection is performed on various classification algorithms like KNN, Decision Tree, SVM, Naive Bayes. Also, after validating if a person has Disease or not, K-Means is applied to perform clustering in 3 clusters of Low, Medium and High Probability of the Disease. Additionally, there is a focus on interpretability, meaning that the system is designed to provide explanations for its predictions, which enhances trustworthiness.

Specifically, for predicting diabetes, heart disease, and kidney disease, the Random Forest algorithm is employed due to its ability to handle complex datasets and provide accurate predictions across multiple classes. On the other hand, the Support Vector Machine algorithm is chosen for predicting Parkinson's disease, given its effectiveness in handling high-dimensional data and identifying complex patterns. Moreover, the system places a strong emphasis on interpretability, providing explanations for its predictions to enhance trustworthiness. Validation against real-world datasets is also emphasized to ensure the system's reliability and effectiveness in practical healthcare settings. Overall the proposed system integrates diverse health data sources, utilizes multiple machine learning algorithms, and emphasizes transparency a validation to revolutionize disease prediction and improve healthcare outcomes.

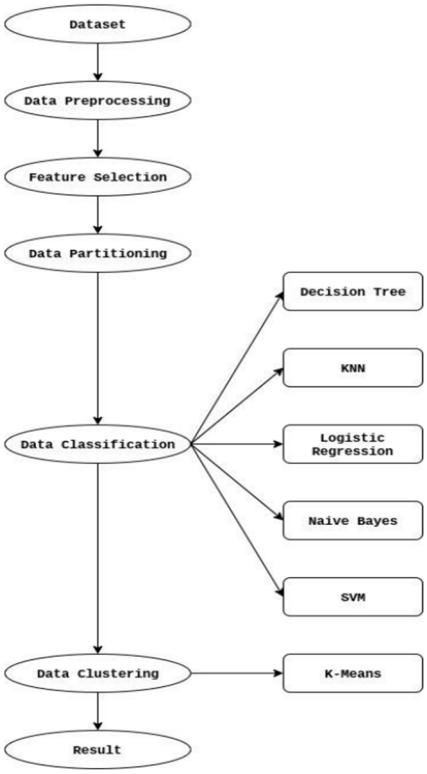


Fig 1 :- Architecture of Disease and Risk Prediction System

IV. PROPOSED METHODOLOGY

1) *Dataset Collection:* This stage involves the collection and comprehension of the dataset to unveil concealed patterns and trends that are crucial for prediction and result evaluation. The dataset comprises a total of 1405 rows, representing the data points, and 10 columns, representing the various features. These features encompass attributes like Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age, Gender, etc.
2) Data Pre-processing: Handle missing values that is Impute missing data or remove incomplete records. Encode categorical variables that is Convert categorical variables into numerical format for model compatibility. Normalize or scale numerical features that is Ensure features are on a similar scale to prevent bias in algorithms.
3) Feature Selection: Utilize techniques such as correlation analysis, feature importance, or domain knowledge to select the most informative features for disease prediction.
4) *Data Partitioning:* The dataset is prepared for training and testing. In the train/split method, the dataset is randomly divided into training and testing sets. Specifically, 1600 samples are allocated for training, while 400 samples are reserved for testing. Optionally, utilize techniques like cross-validation or stratified sampling to ensure robustness and generalizability.
5) *Data Classification:* Implement various machine learning algorithms for disease prediction, including:

a) Decision Tree: Builds a tree-like model based on feature splits.
b) *Random Forest:* Handle complex datasets and provide accurate predictions across multiple classes
c) *K-Nearest Neighbours (KNN):* Classifies instances based on the majority vote of their neighbours.
d) *Logistic Regression:* Models the probability of a binary outcome using a logistic function.
e) *Naive Bayes:* Assumes independence between features and predicts based on Bayes' theorem.
f) *Support Vector Machines (SVM):* Separates data points using a hyperplane to maximize margin.
g) *Gradient Boosting:* Builds an ensemble of weak learners sequentially to improve predictive performance.

6) *Data Clustering:* Apply the K-Means algorithm for data clustering to identify groups of similar data points. Cluster patients based on their features to discover patterns and potential disease subtypes.
7) *Result:* Evaluate the performance of each algorithm using appropriate metrics such as accuracy, precision, recall, and F1-score. Compare the predictive performance of different algorithms to identify the most effective ones for disease prediction. Interpret the results to gain insights into disease patterns, risk factors, and potential interventions.

V. ALGORITHM

1) *Support Vector Machine for Parkinson's Disease:* Using Support Vector Machine (SVM) for Parkinson's disease prediction involves several steps. First, gather a dataset containing relevant features such as patient demographics, medical history, and motor function assessments. Preprocess the data by handling missing values and encoding categorical variables. Then, split the dataset into training and testing sets. Next, train the SVM model on the training data, adjusting parameters like kernel type and regularization strength to optimize performance. Evaluate the model's accuracy, precision, recall, and F1-score on the testing set. Interpret the results to assess the SVM's effectiveness in predicting Parkinson's disease SVM models for Parkinson's disease prediction can achieve accuracy rates ranging from around 80% to 90% in research studies. Additionally, the accuracy of the SVM model may vary depending on the specific characteristics of the dataset and the specific implementation details of the model. Therefore, it's essential to thoroughly evaluate the SVM model's performance on a representative dataset to determine its accuracy for predicting Parkinson's disease accurately.
2) *Random Forest for Diabetes, Heart and Kidney Diseases:* Generally, Random Forest models are known for their robustness and ability to handle complex datasets, making them suitable for disease prediction tasks. In research studies, Random Forest models for predicting heart disease have achieved accuracy rates ranging from approximately 80% to 90%. Similarly, for diabetes prediction, Random Forest models have demonstrated accuracies ranging from around 70% to 85%. For kidney disease prediction, Random Forest models have shown accuracies ranging from approximately 80% to 95%. It's important to note that these accuracy ranges are approximate and can vary depending on the specific characteristics of the dataset and the implementation details of the Random Forest model. Additionally, the accuracy of the model should be evaluated alongside other metrics such as precision, recall, and F1-score to assess its overall performance comprehensively. Therefore, thorough evaluation on representative datasets is crucial to determine the accuracy of a Random Forest model for predicting heart disease, diabetes, and kidney disease accurately.

VI. RESULT AND DISCUSSIONS

After applying Support Vector Machine and Random Forest Algorithm on dataset we got accuracies as mentioned below. Random Forest gives highest accuracy of 79% for diabetes prediction

A. *Accuracy Charts*

Algorithm	Accuracy(%)
Random Forest Classifier	92.54
Decision Tree Classifier	89.47
Gradient Boosting Classifier	89.04
Logistic Regression	88.16
XG Boost	87.72
Support Vector Machine	84.21
K-Nearest Neighbour	83.33

Fig 1. Accuracy of Algorithms for Diabetes Prediction

Algorithm	Accuracy(%)
Random Forest Classifier	0.987013
Decision Tree Classifier	0.977273
Gradient Boosting Classifier	0.977273
Logistic Regression	0.883117
XG Boost	0.977273
Support Vector Machine	0.717532
K-Nearest Neighbour	0.860390

Fig 2. Accuracy of Algorithms for Heart Disease Prediction

Algorithm	Accuracy(%)
Random Forest Classifier	0.991667
Decision Tree Classifier	0.941667
Gradient Boosting Classifier	0.975000
Logistic Regression	0.908333
XG Boost	0.966667
Support Vector Machine	0.700000
K-Nearest Neighbour	0.700000

Fig 3. Accuracy of Algorithms for Kidney Disease Prediction

Algorithm	Accuracy(%)
Random Forest Classifier	0.9661
Decision Tree Classifier	0.9322
Logistic Regression	0.8305
XG Boost	0.9152
Support Vector Machine	0.9661
K-Nearest Neighbour	0.9661

Fig 4. Accuracy of Algorithms for Parkinson's Disease Prediction

B. Graphs

1) Diabetes: Accuracy by Random Forest- 92.54%

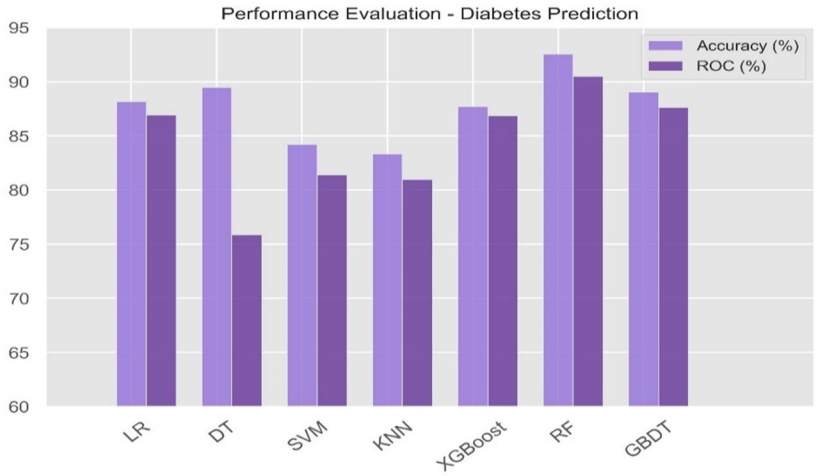


Fig 5. Graphical Representations of the algorithms used for Diabetes Disease Prediction

2) Heart Disease: Accuracy by Random Forest- 98.70%

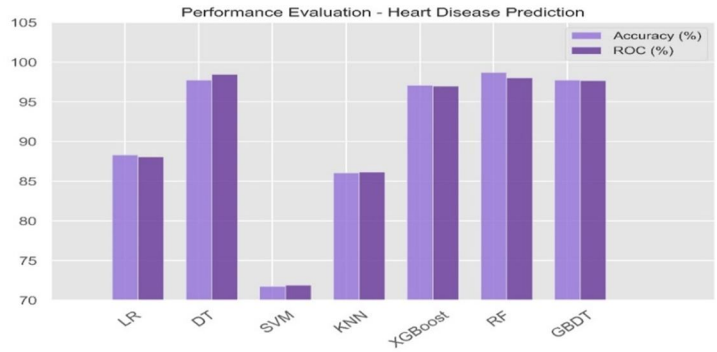


Fig 6. Graphical Representations of the algorithms used for Heart Disease Prediction

3) Kidney Disease: Accuracy by Random Forest- 99.17

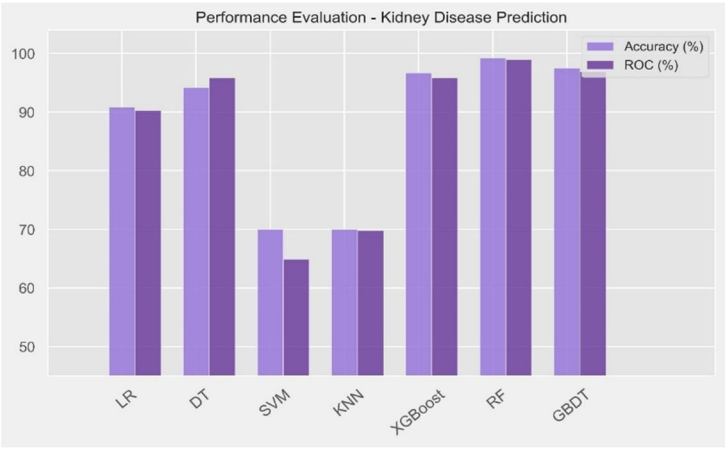


Fig 7. Graphical Representations of the algorithms used for Kidney Disease Prediction

4) Parkinson's Disease: Accuracy by Support Vector Machine- 96.61

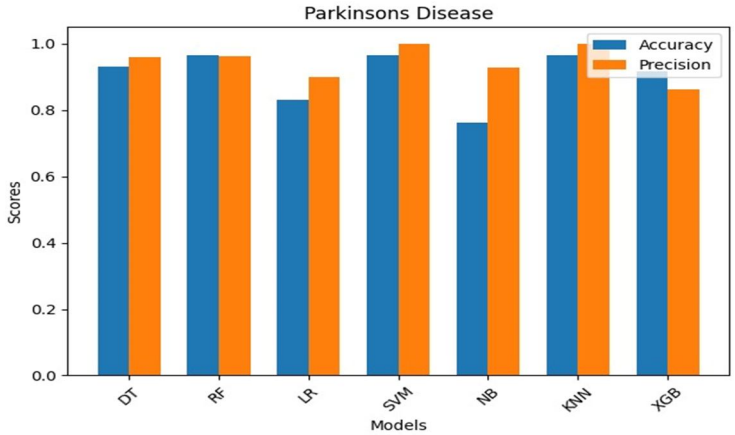


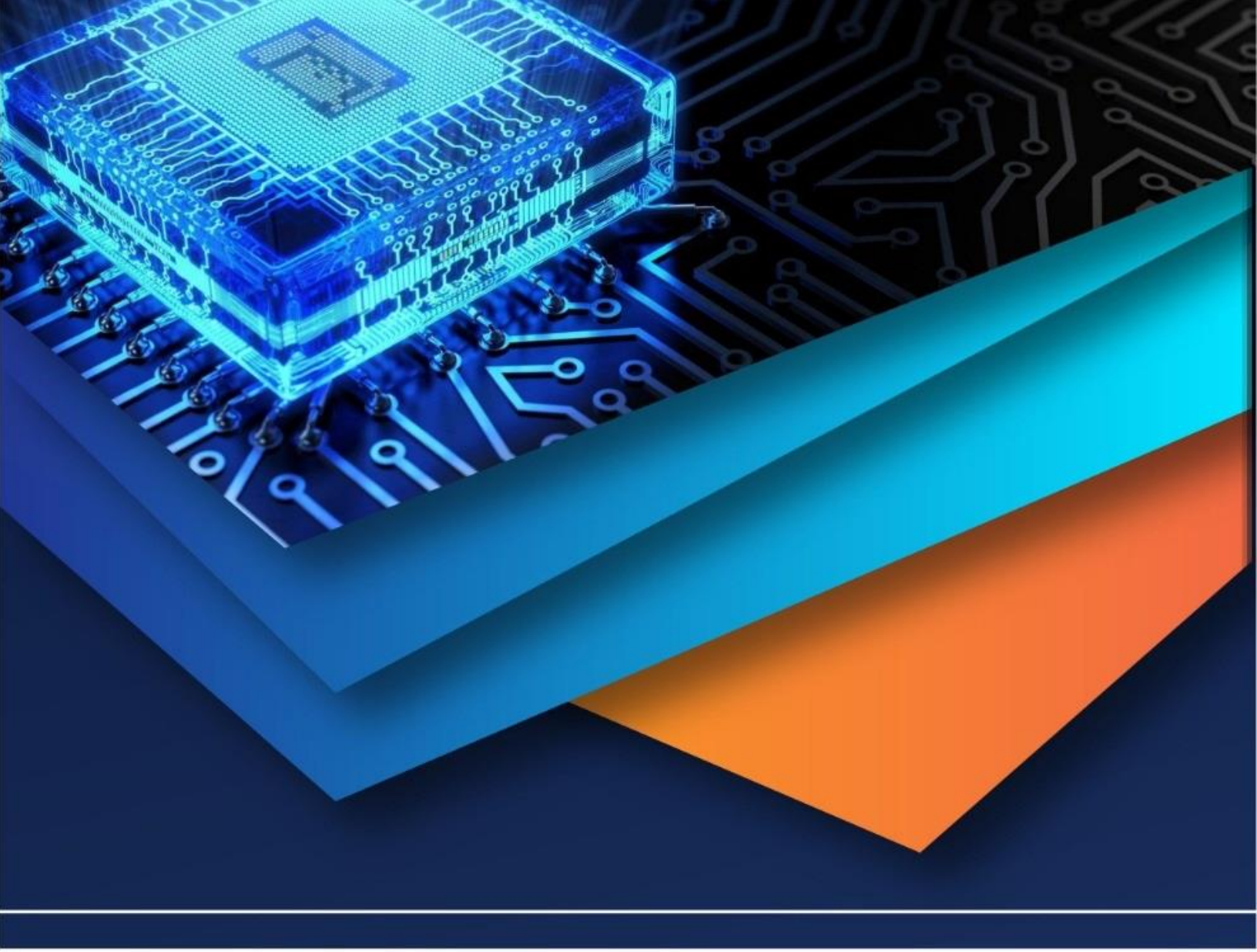
Fig 8. Graphical Representations of the algorithms used for Parkinson's Disease Prediction

VII. CONCLUSION

Detecting diseases at their early stages is indeed a crucial and significant problem in the field of healthcare. In this study, systematic efforts are made in designing a system which results in the prediction of disease. Random Forest outperformed SVM in predicting Diabetes with an accuracy of 92.54%, compared to SVM's 84.21% accuracy. While both models achieved reasonable accuracy, Random Forest showed a slight advantage in this dataset. SVM achieved a higher accuracy of 0.717532% in predicting Heart Disease, while Random Forest achieved an accuracy of 0.987013%. SVM demonstrated superior performance in this dataset, potentially due to the nature of the data and SVM's ability to find complex decision boundaries. Random Forest significantly outperformed SVM in predicting Parkinson's Disease, with an accuracy of 91% compared to SVM's 83% accuracy. Random Forest demonstrated its strength in handling this dataset, which might involve complex relationships between features.

REFERENCES

[1] Priyanka Sonar, Prof. K. JayaMalini," DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING APPROACHES", 2019 IEEE ,3rd International Conference on Computing Methodologies and Communication (ICCMC)
[2] Archana Singh ,Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms", 2020 IEEE, International Conference on Electrical and Electronics Engineering (ICE3)
[3] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques SN Computer Science,1(6).
[4] Katarya, R., & Srinivas, P. (2020, July). Predicting heart disease at early stages using machine learning: a survey. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 302-305). IEEE.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY