STM JOURNALS

Journal of
Computer Technology & Applications

ISSN: 2229-6964 (Online)
ISSN: 2347-7229 (Print)
Volume 16, Issue 2, 2025
May–August
DOI (Journal): 10.37591/JoCTA

https://journals.stmjournals.com/jocta

Review JoCTA

Predicting Multiple Diseases Using Machine Learning: A Data-Driven Approach

Sushma Malik[1], Anamika Rana[2,*], Kanika[3], Ashima[3]

Abstract

The increasing prevalence of chronic and life-threatening diseases highlights the need for innovative healthcare solutions that enable early detection and proactive management. The Multiple Disease Prediction Platform is a web-based system utilizing machine learning (ML) and deep learning (DL) algorithms to analyze user-inputted health data, generating real-time predictions of potential health risks. By leveraging Python's Streamlit library, the platform provides an interactive and accessible diagnostic experience, eliminating the need for frequent clinical visits and enhancing remote healthcare accessibility. This research focuses on developing a supervised learning-based system trained on credible datasets (e.g., Kaggle) for disease prediction. Exploratory and descriptive research methods were employed, incorporating statistical analysis and data visualization to enhance accuracy. Despite its potential, the platform faces limitations, including data privacy concerns, regional biases, reliance on incomplete user data, and computational demands affecting cost and accessibility in resource-limited settings. The study has significant practical implications in advancing digital healthcare, enabling early disease risk assessment and improving decision making for individuals and healthcare professionals. Integrating bio-inspired algorithms further enhances predictive performance by optimizing model efficiency. While deep learning improves complex pattern recognition, ongoing research is essential to refine these models for greater accuracy, scalability, and reliability. Addressing ethical artificial intelligence challenges, model biases, and computational efficiency is crucial for ensuring broad applicability and long-term sustainability.

Keywords: Machine learning, deep learning, disease prediction, digital healthcare, bio-inspired algorithms, supervised learning, data privacy, predictive analytics, health informatics, remote healthcare

INTRODUCTION

In today's world, we are witnessing an alarming surge in diseases, affecting not only the elderly but even the youngest among us. This rise in health crises can be attributed to a myriad of factors—ranging from environmental degradation, changing lifestyles, and increasing stress, to the hidden perils of modern-day pollutants. While these threats may seem inevitable, what is within our control is the power to detect these conditions early and take decisive action. Preventative measures such as regular health check-ups, a balanced lifestyle, and timely interventions can act as our shield against these growing dangers. It is only through vigilance and proactive care that we can hope to combat the seemingly unstoppable tide of illness, safeguarding our minds and bodies for a healthier tomorrow. Taking account of all draconian issues, we are developing a multiple disease prediction system using machine learning algorithms. This minor project focuses on amalgamation of current technology with the day-to-day issues faced by humankind.

***Author for Correspondence**
Anamika Rana
E-mail: anamica.rana@gmail.com

[1]Associate Professor, Department of Computer Applications, Maharaja Surajmal Institute, Janakpuri, Delhi, India
[2]Assistant Professor, Department of Computer Applications, Maharaja Surajmal Institute, Janakpuri, Delhi, India
[3]Scholar, Department of Computer Applications, Maharaja Surajmal Institute, Janakpuri, Delhi, India

Received Date: March 01, 2025
Accepted Date: March 08, 2025
Published Date: April 15, 2025

Citation: Sushma Malik, Anamika Rana, Kanika, Ashima. Predicting Multiple Diseases Using Machine Learning: A Data-Driven Approach. Journal of Computer Technology & Applications. 2025; 16(2): 16–35p.

Machine learning methods have been successfully applied to various fields, including forecasting diseases. The goal of creating a classifier system with machine learning algorithms is to greatly aid in the resolution of health-related problems by helping doctors identify and diagnose illnesses early on [1].

Using a variety of data mining tools and machine learning techniques, one may find important patterns and discover correlations and links between several variables in large databases has altered healthcare institutions [2]. It provides and compares current data for the next course of action, making it a crucial tool in the medical field. This technology makes it possible to explore vast volumes of data by fusing many analytical approaches with sophisticated, contemporary algorithms [3].

A multiple disease prediction model represents a revolutionary step in personalized healthcare. It brings together vast datasets of various diseases, enabling a comprehensive analysis of health risks with remarkable precision. Imagine a platform where, with just a few clicks, users can select the specific illness they are concerned about—whether it is diabetes, heart disease, or any other condition. Leveraging advanced machine learning algorithms and historical medical data, the model swiftly analyzes key health indicators and predicts whether the user may be at risk. The smooth combination of data and technology enables people to take preventive measures for their health, facilitating early diagnosis and prompt treatment. By transforming data into actionable insights, the model offers a powerful tool to combat illness, helping to ensure a future where prevention is as accessible as treatment.

Finally, we propose that the multiple disease prediction model is an innovative combination of machine learning and medicine that makes it possible to diagnose diseases early and with a high degree of accuracy. This technology enables people to take preventative action by evaluating large datasets and detecting health hazards. It marks a major advancement in individualized healthcare, making prevention more widely available, enhancing general health, and lessening the effects of disease.

LITERATURE REVIEW

This section provides an in-depth review of existing research on the application of machine learning in the healthcare industry. By examining previous studies, we tried to draw out their key findings (Table 1). By studying all similar work in this area we aim to highlight the key trends, approaches, and obstacles that have arisen in this swiftly advancing domain.

Table 1. Review of existing research on the application of machine learning in the healthcare industry.

S.N.	Authors	Year	Title	Journal / Book / Conference Name	Methodology	Key findings
1.	Devansh Shah, Samir Patel & Santosh Kumar Bharti	2020	Heart Disease Prediction Using Machine Learning Techniques	Springer Nature Computer Science	Original Research	Demonstrated how data mining and machine learning techniques, utilizing algorithms such as naïve Bayes, decision tree, K-nearest neighbor, and random forest, enable effective heart disease prediction, leading to early diagnosis and improved management in healthcare.

2.	K. Arumugam, Mohd Naved, Priyanka P. Shinde, Orlando Leiva-Chauca, Antonio Huaman-Osorio, Tatiana Gonzales-Yana	2023	Multiple disease prediction using Machine learning algorithms	Materials Today: Proceedings	Case Study	Demonstrated how data mining and machine learning techniques, particularly a fine-tuned decision tree model, enhance prediction accuracy for heart disease in diabetic individuals, aiding in more effective therapy evaluation and management in healthcare.
3.	Mehrbakhsh Nilashi, Othman bin Ibrahim, Hossein Ahmadi, Leila Shahmoradi	2017	An analytical method for diseases prediction using machine learning techniques	Computers & Chemical Engineering	Analytic Study	Demonstrated how a knowledge-based system, combining CART with clustering, noise removal, and fuzzy rule-based techniques, enhances disease prediction accuracy, supporting medical practitioners in clinical decision-making through improved analytical methods.
4	Aishwarya Mujumdar, V Vaidehi Dr.	2019	Diabetes Prediction Using Machine Learning Algorithms	Procedia Computer Science	Case Study and Project	This study demonstrates how a knowledge-based system combining CART, clustering, noise removal, and fuzzy rule-based techniques enhances diabetes prediction accuracy, aiding medical practitioners in clinical decision-making with improved classification and insights from large healthcare datasets.

5	Khader Shameer, Kipp W Johnson, Benjamin S Glicksberg, Joel T Dudley, Partho P Sengupta	2021	Machine Learning in Cardiovascular Medicine: Are We There Yet?	Heart	Case Study	This study outlines how AI techniques—including machine learning, deep learning, and reinforcement learning—enhance cardiovascular medicine by supporting risk prediction, disease phenotyping, and clinical decision-making, while addressing challenges like data heterogeneity, model selection, and method limitations.
6.	Kedar Pingale, Sushant Surwase, Vaibhav Kulkarni, Saurabh Sarage, Prof. Abhijeet Karve	2019	Disease Prediction using Machine Learning	International Research Journal of Engineering and Technology (IRJET)	Case Study & Project	Demonstrated how predictive modeling with Naïve Bayes, linear regression, and decision tree algorithms enhances disease prediction accuracy, enabling early detection and improved patient care across conditions like diabetes, malaria, jaundice, dengue, and tuberculosis.
7.	Mohammed Khalilia, Sounak Chakraborty & Mihail Popescu	2011	Predicting Disease Risks from Highly Imbalanced Data Using Random Forest	BMC Medical Informatics and Decision Making	Comprehensive Study and Project	This study utilizes HCUP's NIS dataset and a random forest-based ensemble approach with repeated random sub-sampling to predict the risk of eight chronic diseases, achieving an average AUC of 88.79%, and providing valuable insights for healthcare decision support.

8.	Y. Deepthi, K. Pavan Kalyan, Mukul Vyas, K. Radhika, D. Kishore Babu & N. V. Krishna Rao	2020	Disease Prediction Based on Symptoms Using Machine Learning	Springer Nature Energy Systems, Drives and Automations: Proceedings of ESDA 2019	Comprehensive Study	This paper explores disease prediction based on symptoms using machine learning algorithms like Naive Bayes, Decision Tree, and Random Forest on healthcare data. Implemented in Python, the study identifies the most accurate algorithm, facilitating efficient diagnosis through big data analytics.
9.	Pahulpreet Singh Kohli; Shriya Arora	2018	Application of Machine Learning in Disease Prediction	IEEE 2018 4th International Conference on Computing Communication and Automation (ICCCA)	Comprehensive Study	This study applies various classification algorithms on three disease datasets (Heart, Breast Cancer, Diabetes) from the UCI repository to aid in early detection. Using backward modeling and p-value testing for feature selection, results highlight machine learning's role in improving diagnosis.
10.	Saraswati Patil, Sangita Jaybhaye, Sujal Bokariya, Pranav Jain, Siddhi Phapale and Tejas Hande	2023	Parkinson's Disease Prediction System in Machine Learning	ITM Web of Conferences	Case Study and Project	Demonstrated how a machine learning-based prediction system, using models like Gradient Boosted Tree, random forest, and logistic regression, enhances early detection of Parkinson's disease, enabling improved disease management and treatment, and providing a scalable framework for diagnosing other neurological disorders.

DATA EXPLORATION

For model execution, we analyzed datasets on various diseases sourced from Kaggle. (diabetes disease prediction, kidney disease, Parkinson's disease and cardiovascular disease). This study was carried out using Python along with various libraries such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn. The study demonstrates how machine learning can be effectively integrated into the healthcare industry, showcasing its potential to enhance disease prediction and early diagnosis. By leveraging data-driven insights, this model aims to improve healthcare accessibility and decision-making.

Problem Formulation

Cardiovascular Disease Prediction

- *Problem statement:* Cardiovascular diseases (CVDs), which include disorders impacting the heart and blood vessels, cause millions of fatalities each year. The primary causes include unhealthy behaviors, environmental risk factors, and comorbidities such as diabetes and hypertension [4].
- *Objective:* Create a model capable of identifying individuals at high risk of developing CVD using a dataset with attributes like age, body mass index (BMI), blood sugar levels, and cholesterol markers. The dataset for the evaluation and development is collected from Kaggle [5].

Diabetes Prediction

- *Problem statement:* Diabetes refers to a range of metabolic conditions characterized by high blood sugar levels, which can result in serious complications like heart disease, kidney failure, and blindness if not properly managed. Timely diagnosis is crucial to avoiding long-term health problems [6].
- *Objective:* Utilize diagnostic measurements from the National Institute of Diabetes and Digestive and Kidney Diseases dataset [7] to predict the diabetes status of patients, focusing on specific features such as glucose levels, BMI, and age.

Parkinson's Disease Prediction

- *Problem statement:* Parkinson's disease is a progressive neurological disorder that significantly impacts an individual's quality of life, causing movement difficulties, tremors, and speech impairments. Timely detection is crucial for more effective disease management [8].
- *Objective:* Employ a Kaggle dataset [9] containing biomedical voice measurements to classify individuals as healthy or Parkinson's patients, based on features such as frequency variations and voice amplitude.

Chronic Kidney Disease Prediction

- *Problem statement:* Chronic kidney disease (CKD) is a life-threatening condition that requires timely intervention to prevent progression to end-stage renal failure. Risk factors like high blood pressure, diabetes, and genetic factors elevate the likelihood [10].
- *Objective:* Develop a predictive model using a dataset from the University of California Irvine (UCI) Machine Learning Repository, which includes attributes like blood urea levels, creatinine levels, and kidney functioning indicators. The dataset is collected from archive.ics.uciedu [11].

Data Cleaning

There are various standardized techniques for data preprocessing, including methods like data cleaning, normalization, and encoding, all of which ensure the dataset's consistency and precision. Effective preprocessing is vital for optimizing the performance of predictive models and algorithms. In this study, we ensured the proper data type for each column, such as ensuring the date column was in the correct date and time format for accurate analysis. We also identified and handled outliers and missing or null values by removing inconsistencies. These steps were essential in refining the dataset for further testing and analysis.

Predictor Selection

We have carefully analyzed the datasets and selected the relevant features to be included in the model. This process ensures that only the important columns are used for analysis, thereby avoiding the curse of dimensionality. Feature selection (FS) techniques are commonly applied in data preprocessing to reduce data dimensions effectively, aiding in the discovery of more accurate models. As a comprehensive search for the best feature subset is often impractical, various search strategies have been introduced in the literature. FS is typically used in tasks such as classification, clustering, and regression to improve model performance [12].

Exploratory Data Analysis

In this section, we conducted a thorough analysis of our datasets, focusing on their attributes. We explored the statistical relationships between these attributes to gain a deeper understanding of how they interact and influence each other. This analysis was crucial for preparing the data for further modeling and interpretation.

Cardiovascular Disease

Diseases that impact the heart or blood vessels are commonly known as cardiovascular disease (CVD). It is typically linked to atherosclerosis, or the accumulation of fatty deposits inside the arteries, and a higher risk of blood clots. It can also be associated with damage to the arteries in organs such as the brain, heart, kidneys, and eyes [13]. We used a very thorough dataset from Kaggle [14], which included many important features, to develop a prediction model for cardiovascular illness.

Given the presence of noise within the dataset, we meticulously addressed this issue by eliminating outliers and establishing rigorous criteria, thereby ensuring data consistency for robust exploratory data analysis and predictive modeling.

The accompanying graphs, which provide a clear and thorough comprehension of important patterns and trends, graphically depict the important insights we have discovered after a thorough and painstaking investigation of the information.

The cardiovascular heatmap illustrates correlations between different health indicators. It is important to highlight that there is a significant positive relationship between systolic (ap_hi) and diastolic (ap_lo) blood pressure, with a correlation of 0.71., suggesting their close relationship in cardiovascular risk assessment. Cholesterol and glucose levels have a moderate correlation (0.44), indicating a possible interdependence affecting cardiovascular health. Age shows a moderate positive correlation with cardiovascular disease (0.23), emphasizing the increased risk as age progresses. Gender has a notable correlation with smoking habits (0.34), reflecting behavioral trends that may influence health outcomes. Overall, these insights help identify key factors impacting cardiovascular conditions as shown in Figure 1.

The bar chart in Figure 2 depicts the distribution of cardiovascular disease (cardio) status across genders. Gender 1 has a higher count of individuals both without (cardio = 0) and with (cardio = 1) cardiovascular disease compared to Gender 2. This suggests that Gender 1 may be more prone to or represented in cases of cardiovascular issues. The difference in counts indicates a potential gender-based disparity in cardiovascular health that warrants further investigation as shown in Figure 2.

Diabetes

Diabetes encompasses a range of metabolic disorders characterized by elevated blood glucose levels persisting over an extended period. Common symptoms of elevated blood glucose levels include frequent urination, intense thirst, and an increased appetite. Without proper management, diabetes can lead to numerous complications. Immediate complications can include diabetic ketoacidosis, hyperosmolar hyperglycemic state, or even death, whereas serious long-term consequences encompass heart disease, stroke, chronic kidney failure, foot ulcers, and vision loss [15].

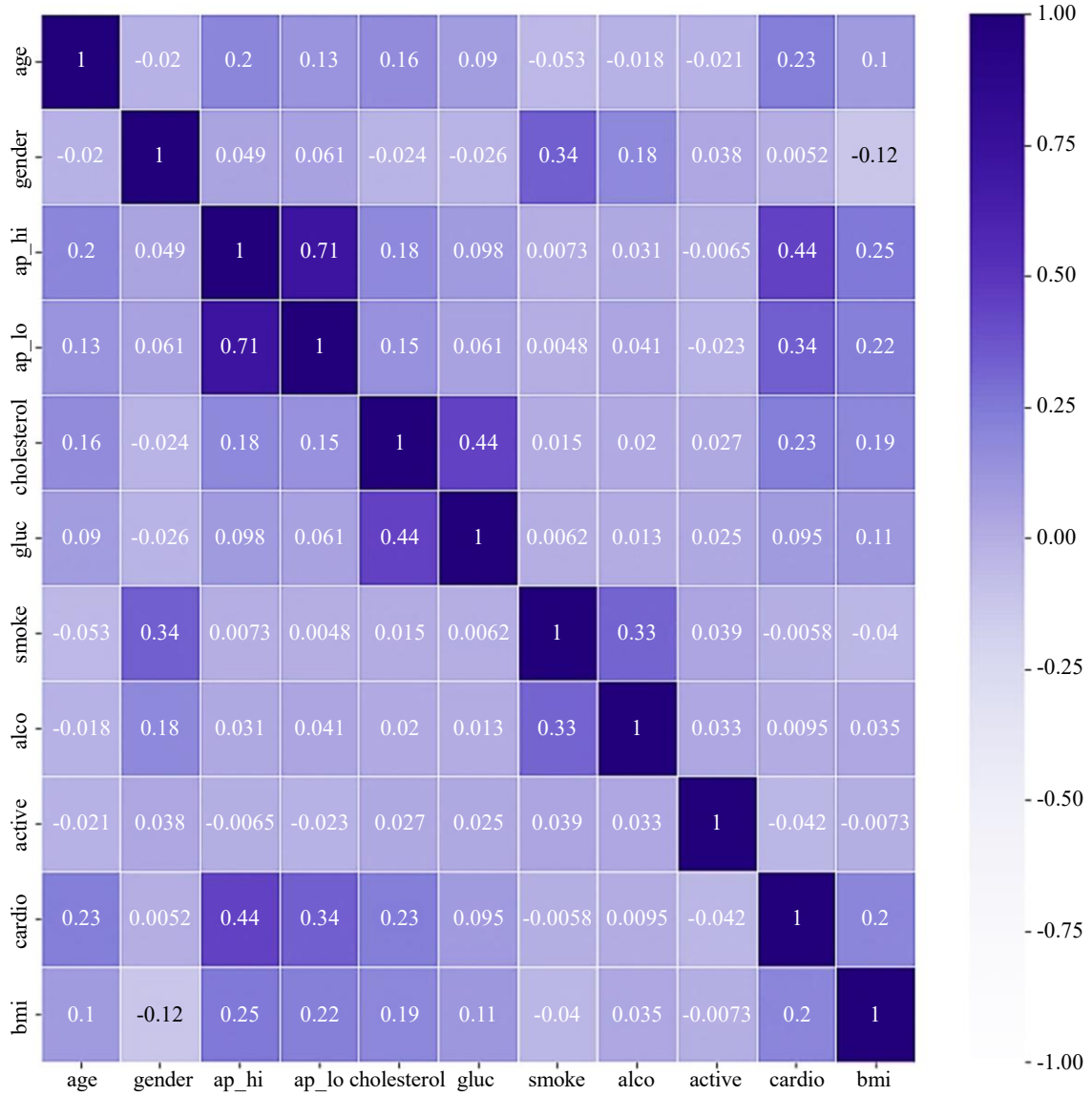

Figure 1. Heat map for the correlation of variables of dataset.

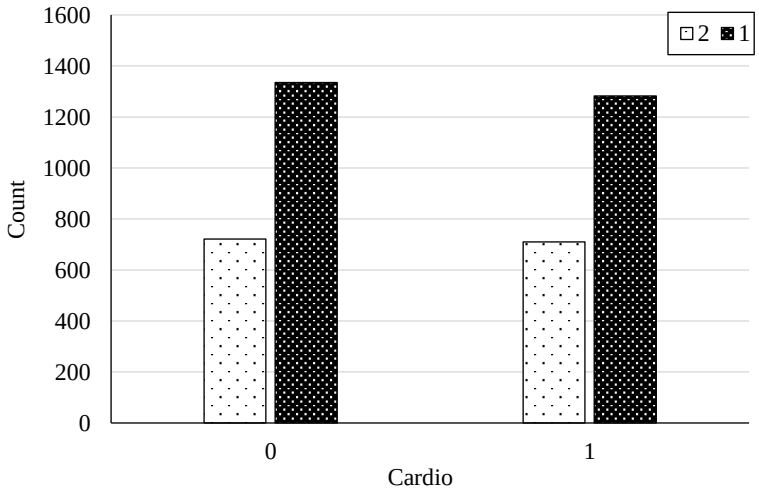

Figure 2. Bar chart showing the count for each gender.

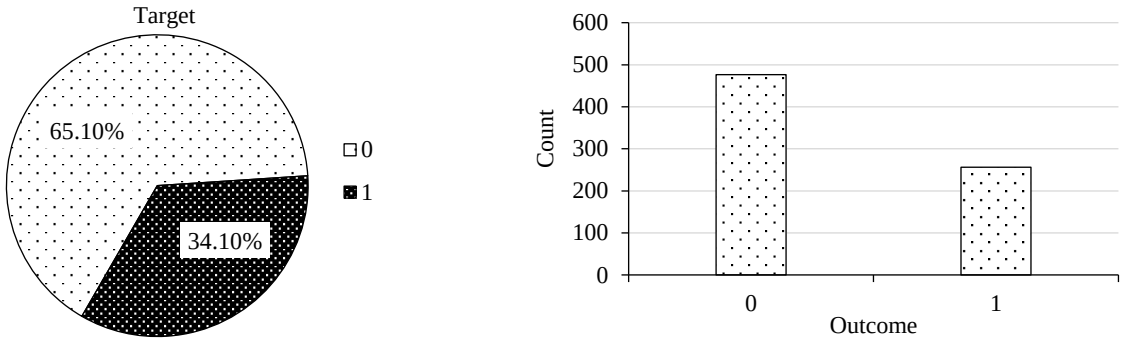

Figure 3. Pie chart shows proportions; count plot shows numbers.

This dataset, sourced from Kaggle [16], aims to predict diabetes status in patients based on specific diagnostic metrics provided within the dataset. This dataset was sourced from the National Institute of Diabetes and Digestive and Kidney Diseases. To improve data consistency and modeling accuracy, we removed outliers and applied strict criteria. Our analysis identified key patterns and trends, which are clearly illustrated in accompanying visual graphs.

The charts in Figure 3 provide insights into the distribution of outcomes within the diabetes dataset. The pie chart on the left shows that 65.1% of individuals do not have diabetes (target = 0), while 34.9% have diabetes (target = 1). The bar chart on the right further reinforces this, displaying a higher count of non-diabetic individuals (outcome = 0) compared to diabetic ones (outcome = 1). This imbalance in the dataset suggests that diabetes cases are fewer, highlighting the need for careful handling in predictive modeling to avoid biased results.

Parkinson's Disease

Parkinson's disease (PD) is a degenerative neurological condition that impacts movement and deteriorates as time progresses. It often begins with subtle symptoms, like a slight tremor in the hand, foot, or jaw. In addition to tremors, the disease also leads to rigidity, reduced movement speed, and balance problems, raising the likelihood of falls. Early signs may include reduced facial expression, a lack of arm swing when walking, and softer or slurred speech. Although there is no cure at present, medications can aid in controlling symptoms and enhancing quality of life [17]. Early detection and diagnosis of PD are vital for effective treatment and management [18].

This dataset, obtained from Kaggle [19], includes a variety of biomedical voice measurements from 31 individuals, 23 of whom have Parkinson's disease (PD). Each column represents a specific voice measurement, and each row corresponds to one of 195 voice recordings from these participants ("name" column). The primary objective of the data is to distinguish between healthy individuals and those with PD, as indicated by the "status" column, where 0 denotes healthy and 1 denotes PD.

The following figures provide a visual summary of critical insights, effectively illustrating the significant patterns and trends identified through an in-depth, meticulous analysis. They offer a clear and thorough understanding of essential information uncovered during the investigation.

The heatmap in Figure 4 presents the correlation among vocal features in PD data, highlighting key relationships. Strong positive (red) or negative (blue) correlations are observed between features like jitter, shimmer, and frequency measures. High inter-correlation among shimmer features suggests they vary together, while varying correlations with "status" (indicating PD) imply that certain vocal features might be more indicative of disease severity, potentially aiding in diagnosis and monitoring.

After applying feature selection to the Parkinson's disease dataset, "Spread1" and "PPE" emerge as the most influential predictors, with high percentile scores indicating their importance in distinguishing Parkinson's presence or severity. Other moderately important features include "D2" and "MDVP:

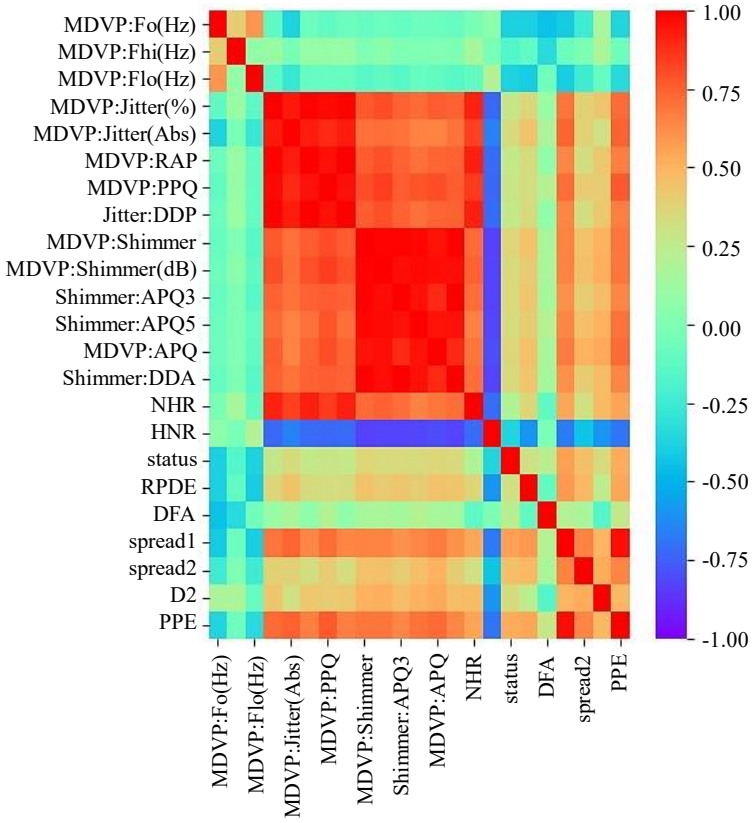

Figure 4. The heatmap of the correlation among vocal features in Parkinson's disease data.

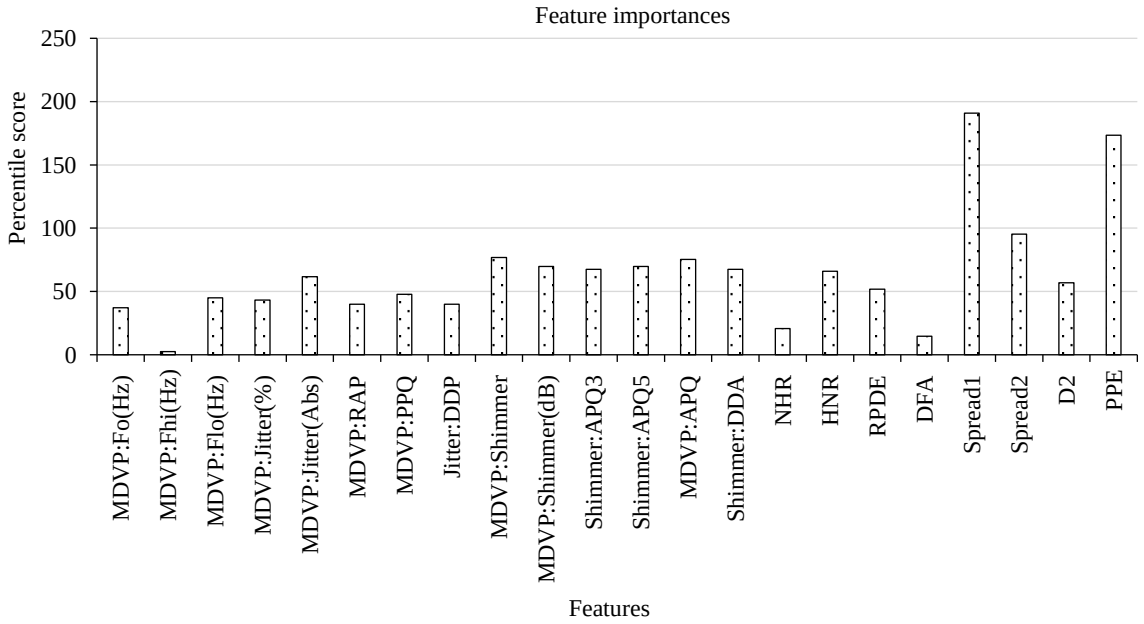

Figure 5. Importance of features.

Shimmer (dB),” suggesting that variability in vocal shimmer and other spread measures contribute to predictive accuracy. Lesser importance is assigned to features like “NHR” and “RPDE,” which may add limited value in this context. This ranking helps prioritize features for modeling PD diagnosis or progression. The chart in Figure 5 depicts the same.

Chronic Kidney Disease

Chronic kidney disease (CKD) is a serious condition that can persist throughout a person's life, resulting from either kidney cancer or reduced kidney function. It is possible to prevent or slow the advancement of this long-term disease to an end-stage, where dialysis or surgery is the only option to sustain the patient's life. Early diagnosis and proper treatment can improve the chances of this occurring. [20]. This dataset is sourced from the UC Irvine Machine Learning Repository [21].

After we have performed the exploratory data analysis (EDA) on the dataset we have made some visualizations. The graphs visually summarize key insights, effectively highlighting significant patterns and trends revealed through detailed analysis. The correlation heatmap in Figure 6 shows the relationships between various factors and kidney disease classification. Darker colors near 1 or -1 indicate strong positive or negative correlations, while lighter colors suggest weaker associations. For example, “hemo” (hemoglobin) and “htn” (hypertension) exhibit moderate correlations with “classification,” indicating relevance to kidney disease status. Other features like “age” and “bu” (blood urea) also show correlations, providing insights into which factors may be significant for disease prediction.

The bar chart in Figure 7 illustrates the distribution of classifications in the dataset, with “0” and “1” representing two different classes, likely indicating the absence or presence of kidney disease. The count for class “1” is higher, suggesting more cases fall under this classification. This imbalance may indicate a greater prevalence of kidney disease in the dataset, which could impact model training and highlight the importance of accounting for class imbalance in analysis.

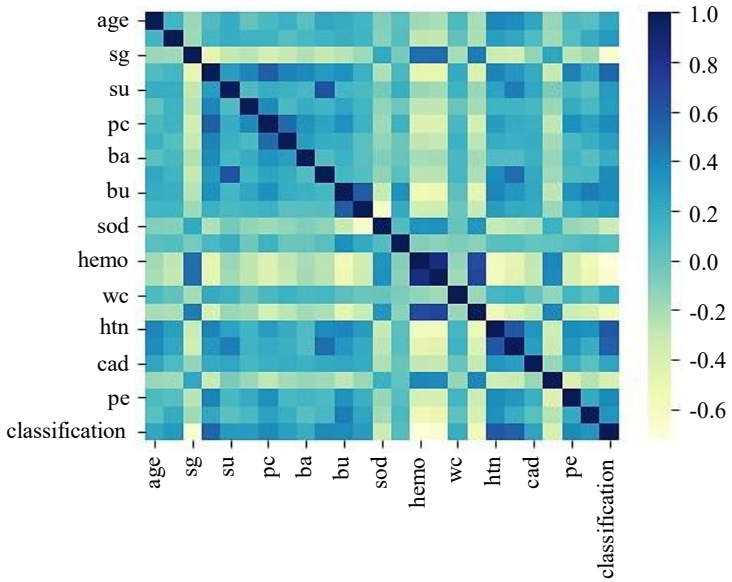

Figure 6. Heat map for the correlation of variables of dataset.

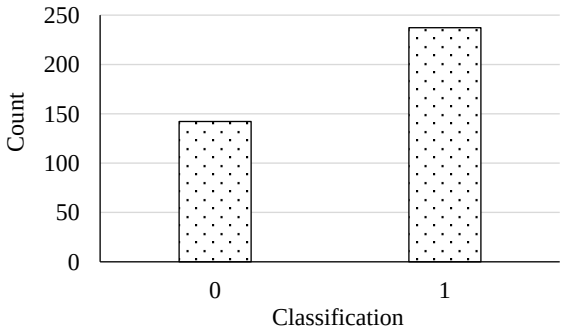

Figure 7. Bar chart illustrating the distribution of classifications.

METHODOLOGY

This section explores the machine learning algorithms used in developing our model. Each disease required a tailored approach, utilizing different algorithms to ensure optimal predictive accuracy. By selecting the most suitable models for each condition, we aimed to enhance diagnostic precision and improve the platform's overall performance in disease prediction.

Random Forest

A random forest classifier is a proven ensemble method that is commonly utilized in machine learning and data science for various applications. This approach employs "parallel ensembling," where multiple decision trees are built simultaneously on various sub-samples of the data, as illustrated in Figure 8.

The model aggregates results through majority voting or averaging, which reduces the risk of overfitting and enhances both prediction accuracy and stability. Consequently, a random forest with multiple trees is generally more precise than a model based on a single decision tree. The technique creates a series of decision trees by combining bootstrap aggregation (bagging) and random selection of features, allowing controlled variation across the trees. It is proficient in both classification and regression tasks, effectively managing both categorical and continuous data [22].

Logistic Regression

Logistic regression (LR) is a probabilistic and a statistical model often used for classification tasks in machine learning [23]. It relies on the logistic, or sigmoid, function to estimate probabilities, helping to assign data points to specific categories. While effective on linearly separable datasets, logistic regression may struggle with high-dimensional data, where it can risk overfitting. To address this, regularization techniques such as L1 and L2 are commonly applied, providing a safeguard against overfitting by constraining the model's complexity [24]. However, a notable limitation of logistic regression is its assumption of a linear relationship between the dependent and independent variables, which may reduce its effectiveness in cases where the data has a non-linear structure. Though logistic regression can be applied to both classification and regression problems, it is more frequently utilized in classification tasks due to its proficiency in estimating discrete outcomes [25].

Sigmoid function (Figure 9):

$g(z) = \frac{1}{1+\exp(-z)}$.

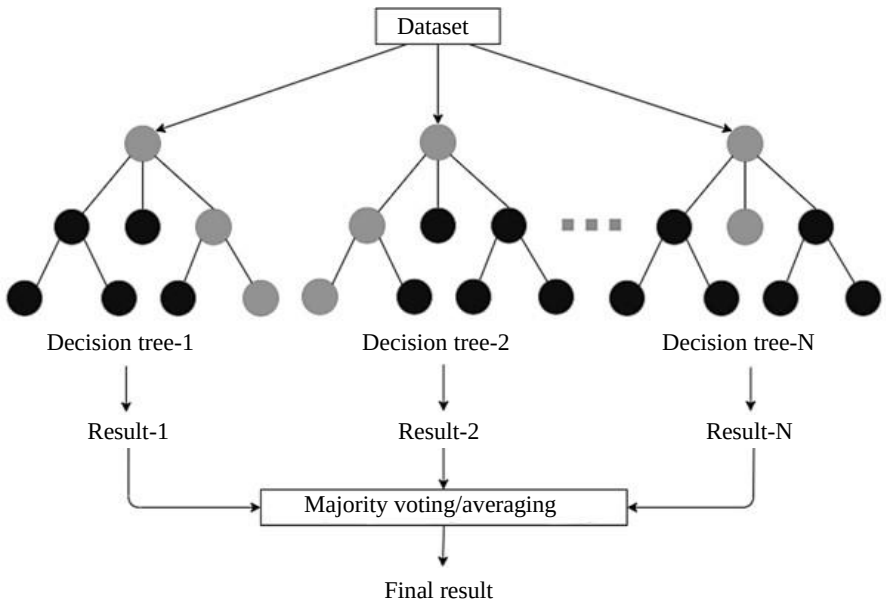

Figure 8. Random forest.

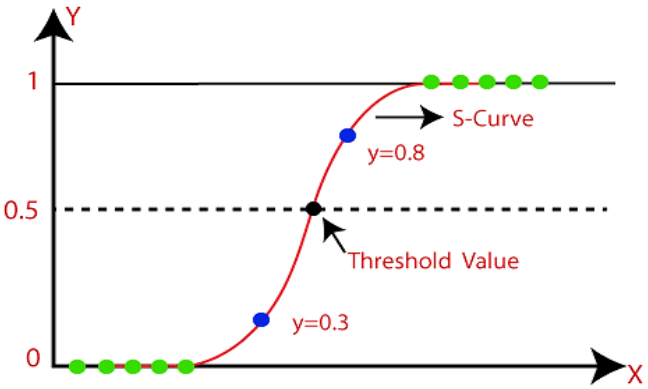

Figure 9. Sigmoid function.

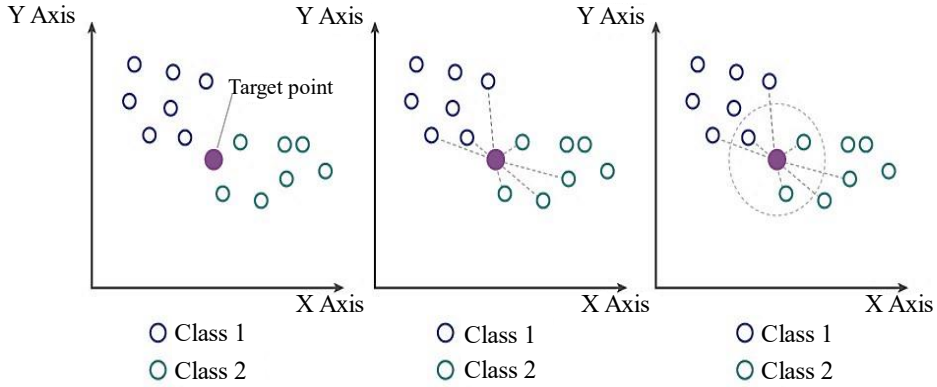

Figure 10. K-nearest neighbor (KNN).

K-Nearest Neighbor

The K-nearest neighbor (kNN) (Figure 10) classification is an advanced method that identifies a set of *k* nearest items from the training data relative to a test item, then assigns a label to the test item based on the majority class within that set. This approach is effective because it acknowledges that exact matches between items in a dataset are rare and that items closest to a test point might provide conflicting class information. Key aspects of kNN include (i) the labeled dataset for determining the test item's class, (ii) a measure of distance or similarity to gauge how close items are, (iii) the selection of *k*, the count of nearest neighbors, and (iv) a rule for classifying the test item based on the classes and distances of these neighbors. The simplest version assigns a class based on either the nearest single neighbor or the majority among the *k* neighbors, though enhancements to this approach are common.

KNN is an instance-based learning method, specifically a subset that includes approaches like case-based reasoning, which focuses on symbolic data. It is also an example of lazy learning, meaning it only generalizes from the training data once a query is made, rather than building a generalized model in advance [26].

Support Vector Machines

Support vector machines (SVMs) (Figure 11) are supervised learning algorithms used for classification by finding the best hyperplane that maximizes the margin between different data classes. Introduced by Vladimir N. Vapnik in 1995, SVMs improve generalization by selecting the hyperplane with the largest margin, enhancing predictive accuracy. They efficiently manage both linear and nonlinear classification problems by utilizing kernel functions like linear, polynomial, radial basis function (RBF), and sigmoid. When data cannot be separated linearly, the "kernel trick" transforms it into a higher-dimensional space, allowing for improved separation and enhanced model performance across different applications [27].

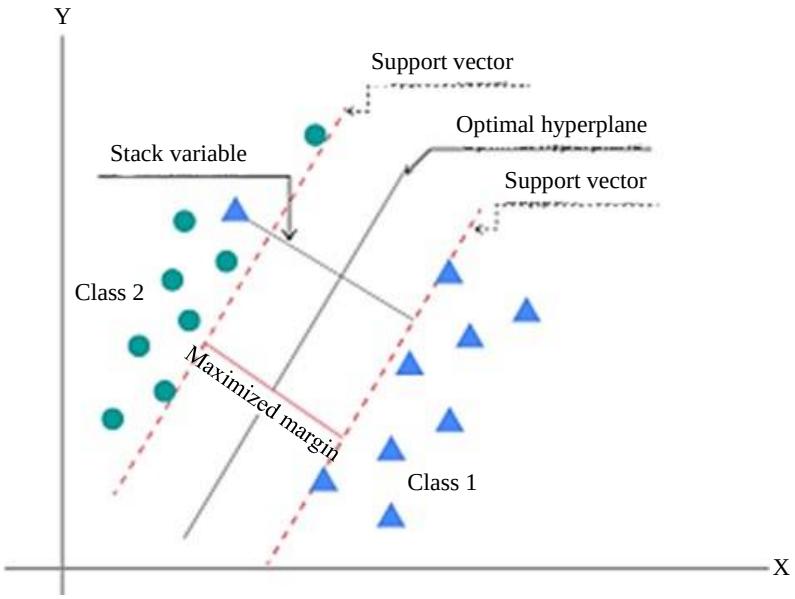

Figure 11. Diagram for support vector machines (SVMS).

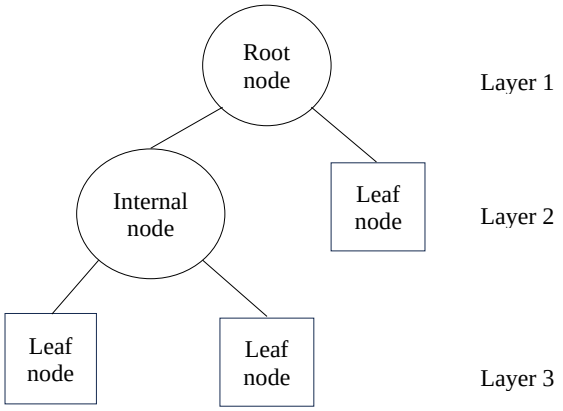

Figure 12. Basic structure of a decision tree [26].

Decision Tree

A decision tree as shown in Figure 12 is constructed from a set of training data using the "divide and conquer" strategy. If all items in the training set belong to the same decision class (i.e., have the same output value), the tree will consist of a single leaf node with that decision label. However, if items belong to multiple decision classes, an attribute that separates them is chosen, dividing the set of items based on the attribute's categories. This chosen attribute forms a test node in the decision tree. The process then recursively applies to each branch stemming from this node, using only the remaining items in each subgroup, until each branch leads to a leaf node representing a decision [28].

In summary, a decision tree is a graphical representation of decisions, and their outcomes structured as a tree. The nodes in the tree represent events or choices, while the edges symbolize the decision rules or conditions. Each tree is made up of nodes and branches, where each node signifies an attribute within a group to be classified, and each branch indicates a possible value that the node can assume [29].

Extreme Gradient Boosting (XGBOOST)

Gradient boosting is a powerful ensemble learning method that creates a robust predictive model by combining a sequence of individual models, most commonly decision trees. In gradient boosting, every model in the sequence is designed to address the mistakes of the previous one, leading to an improved model with higher accuracy. The gradient, representing the direction and rate of change, is applied to

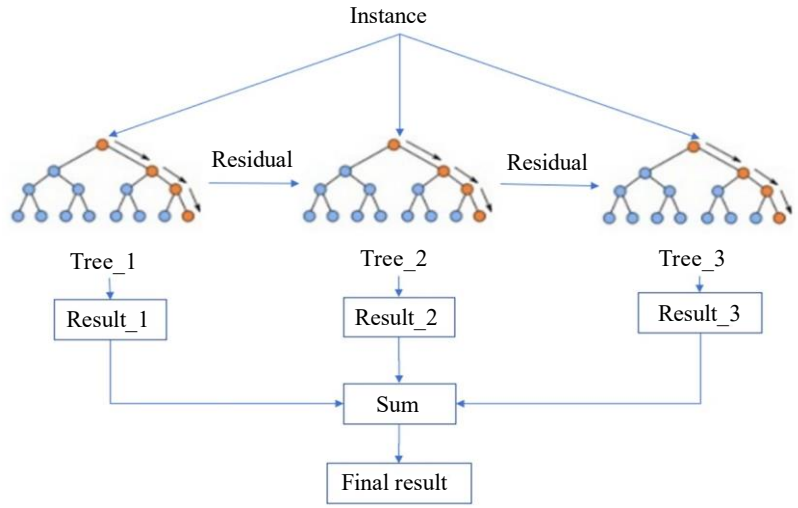

Figure 13. Extreme gradient boosting (XGBoost).

minimize the model's loss function, a method similar to gradient descent in neural networks, where weights are adjusted to optimize performance. A popular variant of gradient boosting is extreme gradient boosting, or XGBoost, which has gained widespread adoption due to its efficiency and effectiveness in handling large datasets. XGBoost enhances traditional gradient boosting by incorporating second-order gradients, which allow it to capture more complex relationships in the data, leading to a more accurate model. Additionally, it uses advanced regularization techniques (L1 and L2) that help reduce overfitting, thus improving the model's ability to generalize well to new, unseen data as shown in Figure 13.

XGBoost stands out for its computational efficiency; it is not only fast to train and interpret but also capable of handling massive datasets without compromising performance. With these advanced features, XGBoost has become a favored tool in data science, offering a powerful combination of precision, scalability, and resilience against overfitting, making it a versatile choice for a range of complex predictive tasks.

AdaBoosting

AdaBoosting operates by initially training a model on a given dataset. If the model makes errors, a new model is created to correct them. This iterative process continues, refining the predictions until errors are minimized. The method functions by merging several weak learners—models that perform marginally better than random guessing—into one powerful learner that can make precise predictions. By adjusting the weights of misclassified data points in each iteration, AdaBoost ensures that subsequent models focus more on difficult cases, ultimately improving overall performance and accuracy. This boosting technique is widely used to enhance classification and predictive modeling tasks [30].

RESULTS AND DISCUSSION

The effectiveness of the Multiple Disease Prediction Model was evaluated using different machine learning algorithms, with a focus on key metrics like accuracy, precision, recall, and F1-score. The analysis focuses on how effectively the model identifies health risks and differentiates between conditions based on input parameters.

Cardiovascular Disease

The bar chart in Figure 14 and Table 2 compare the accuracy scores of three machine learning algorithms: logistic regression, K-nearest neighbors, and random forest. Logistic regression and random forest achieved the highest accuracy scores (both around 74.5%), indicating strong performance on the dataset. KNN, on the other hand, had a lower accuracy score (approximately 71.5%), suggesting it may not generalize as well for this problem.

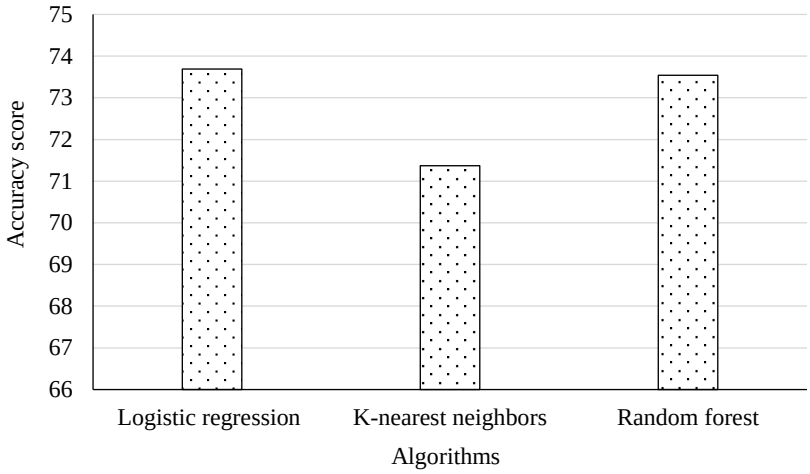

Figure 14. Graphical representation of three machine learning algorithms.

Table 2. Comparative analysis of algorithms on dataset with corresponding accuracy scores

	Algorithms	Score_train	Score_test
0	Logistic regression	0.723333	0.745686
1	K-nearest neighbors	0.753030	0.716621
2	Random forest	0.805758	0.743869

While logistic regression provides consistent and interpretable results, random forest's ability to handle both categorical and numerical data, along with its robustness to overfitting, makes it a strong candidate for model prediction. Given its balance between performance and generalizability, we will proceed with the random forest model for final predictions.

Diabetes

The bar chart in Figure 15 and Table 3 compare the accuracy scores of three machine learning models: random forest (RF), support vector machine (SVM), and XGBoost (XGB). Among them, XGBoost achieved the highest accuracy (approximately 88.9%), indicating its strong predictive power and ability to handle complex patterns. Random Forest came in second with an accuracy of approximately 88.5%. showcasing its reliability and robustness in diverse datasets. SVM had the lowest accuracy (approximately 85.8%), suggesting it may not generalize as well as the other models for this dataset. Given XGBoost's superior accuracy and efficiency in handling both structured and unstructured data, we will proceed with it for final predictions.

Parkinson's Disease

The bar chart in Figure 16 and Table 4 present a comparison of accuracy scores for three machine learning models: support vector machine (SVM), decision tree (DT), and random forest classifier (RFC). Among these models, RFC achieved the highest accuracy (around 91%), demonstrating its strong performance and ability to handle complex data structures. DT followed with an accuracy of approximately 87%, indicating its capability to capture patterns but with potential overfitting. SVM had the lowest accuracy (around 79%), suggesting it may not be the best fit for this dataset. Given its superior accuracy and robustness in handling both categorical and numerical data, we will proceed with the random forest model for final predictions.

Chronic Kidney Disease

The bar chart in Figure 17 compares the accuracy scores of three machine learning models: random forest classifier, AdaBoost classifier, and gradient boosting classifier. The accuracy score is also mentioned in Table 5.

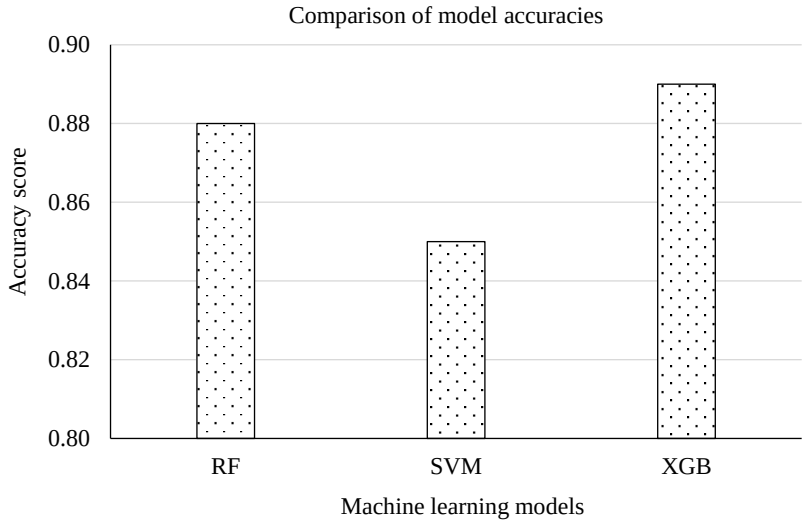

Figure 15. The accuracy scores of three machine learning models.

Table 3. The accuracy levels for all algorithms applied on the dataset.

	Model	Mean accuracy	Standard deviation
0	Random forest (RF)	0.882895	0.039997
1	Support vector machine (SVM)	0.857895	0.036654
2	Extreme gradient boost (XGB)	0.886842	0.027474

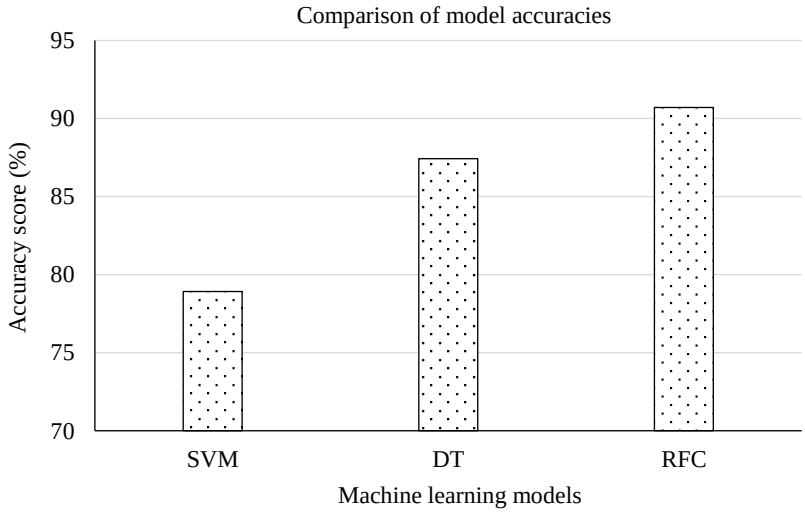

Figure 16. Comparison of accuracy scores for three machine learning models.

Table 4. The accuracy levels for all algorithms applied on the dataset.

	Model	Accuracy score (%)
0	Support vector machine classifier	78.81
1	Decision tree classifier	87.29
2	Random forest classifier	90.68

The random forest and gradient boosting classifiers attained the highest accuracy, reaching about 97.5%. while AdaBoost performed slightly lower (approximately 96.6%). This indicates that both random forest and gradient boosting classifiers are well-suited for this dataset, likely due to their ability to handle complex patterns and interactions between features.

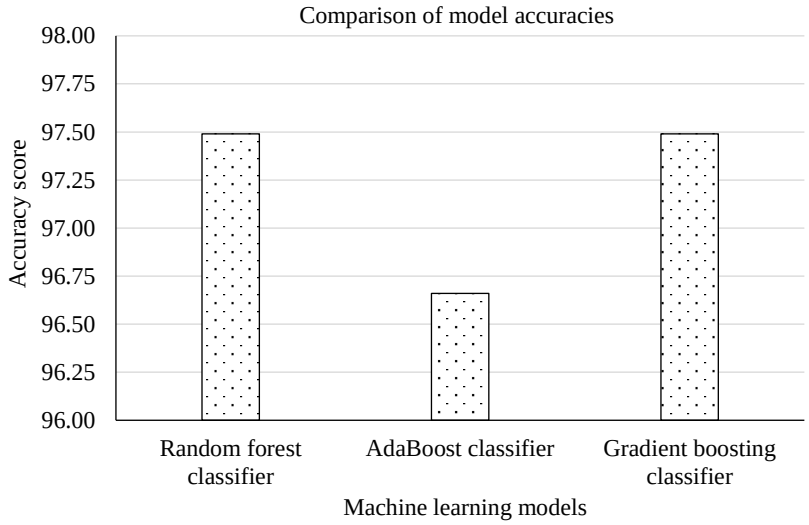

Figure 17. Comparison of model accuracies.

Table 5. The accuracy levels for all algorithms applied on the dataset.

	Model	Accuracy score (%)
0	Random forest classifier	97.50
1	AdaBoost classifier	96.67
2	Gradient boosting classifier	97.50

While AdaBoost showed a slightly lower accuracy, it may still be useful in situations where interpretability and computational efficiency are more important. Given the strong performance of both random forest and gradient boosting, the final model selection will depend on additional factors such as training time, feature importance analysis, and overfitting risk. However, since both random forest and gradient boosting handle numerical and categorical data effectively, we will proceed with one of them for model prediction.

FUTURE IMPLICATIONS

The Multiple Disease Prediction Web App represents a major step toward integrating machine learning into personalized healthcare. As technology advances, the platform's capabilities can be expanded to offer even more comprehensive diagnostic and preventive solutions. One of the most promising future directions is incorporating deep learning techniques, such as convolutional neural networks (CNNs), to analyze complex medical images like X-rays, magnetic resonance imaging (MRI), and compuited tomography (CT) scans. This enhancement would allow the platform to move beyond text-based data and provide a more holistic assessment of a user’s health status.

Another significant improvement lies in integrating Google’s API (application programming interface) tools to provide users with personalized health recommendations based on predictive results. By analyzing a user's health data, the system could suggest preventive measures, lifestyle modifications, or routine screenings, empowering individuals to take proactive steps in managing their health. This feature would make healthcare more accessible by guiding users toward actionable decisions without the need for frequent clinical visits.

Furthermore, improving the quality and diversity of training data will be crucial for enhancing the platform’s predictive accuracy. Instead of relying solely on publicly available datasets, future iterations could incorporate real-world hospital-sourced data. This would ensure that predictions are based on clinically verified health records, making the model more robust and applicable across different demographics. Hospital data would also allow for a more precise understanding of disease patterns, thereby increasing the model’s reliability in medical decision making.

Lastly, as artificial intelligence continues to evolve, integrating natural language processing (NLP) could enhance user interaction. Voice-assisted inputs and chatbot-driven consultations could make the platform even more user-friendly, catering to a wider audience, including elderly individuals and those with limited technical proficiency.

By continuously refining its models and expanding its functionalities, this platform has the potential to become a game-changing tool in predictive healthcare, ultimately improving patient outcomes and reducing the burden on healthcare systems worldwide.

LIMITATIONS

Despite its potential, the Multiple Disease Prediction Web App has some limitations. One primary concern is data privacy and security; using hospital-sourced data necessitates strict adherence to regulations like HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation), which could increase operational complexity and costs. Additionally, model accuracy may vary across diverse populations, as medical data often reflects regional biases, potentially limiting its applicability on a broader scale. Another challenge is that the platform relies heavily on user-input data, which may not always be accurate or complete, affecting prediction reliability. Lastly, integrating advanced features like image analysis via CNNs requires significant computational resources, which may increase deployment costs and impact the platform's accessibility in resource-limited areas. These limitations highlight areas for ongoing improvement and consideration as the platform evolves.

CONCLUSION

In conclusion, disease prediction through machine learning, deep learning, and bio-inspired algorithms holds transformative potential for healthcare. By leveraging diverse algorithms, from K-nearest neighbors and random forest to convolutional neural networks and crow search algorithms, these approaches have demonstrated impressive capabilities in identifying patterns and predicting disease risks accurately. Machine learning provides a structured approach for analyzing complex datasets, while deep learning excels in interpreting images and sequences, enabling early and precise diagnosis. Bio-inspired algorithms bring unique optimization strategies, enhancing prediction performance in various scenarios. Together, these techniques could reshape preventative care, helping medical professionals address health risks proactively. Nevertheless, continuous research is essential to refine these models for even higher accuracy, scalability, and accessibility, ensuring they can serve as practical, reliable tools for the medical community.

REFERENCES

1. Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning techniques. SN Computer Sci. 2020; 1 (6): 345.
2. Arumugam K, Naved M, Shinde PP, Leiva-Chauca O, Huaman-Osorio A, Gonzales-Yanac T. Multiple disease prediction using machine learning algorithms. Mater Today Proc. 2023; 80: 3682–3685.
3. Nilashi M, bin Ibrahim O, Ahmadi H, Shahmoradi L. An analytical method for diseases prediction using machine learning techniques. Computers Chem Eng. 2017; 106: 212–223.
4. Mujumdar A, Vaidehi V. Diabetes prediction using machine learning algorithms. Procedia Computer Sci. 2019; 165: 292–299.
5. Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Machine learning in cardiovascular medicine: are we there yet? Heart. 2018; 104 (14): 1156–1164.
6. Pingale K, Surwase S, Kulkarni V, Sarage S, Karve A. Disease prediction using machine learning. Int R J Eng Technol. 2019; 6 (12): 831–833.
7. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Med Informatics Decis Making. 2011; 11: 1–3.
8. Deepthi Y, Kalyan KP, Vyas M, Radhika K, Babu DK, Krishna Rao NV. Disease prediction based on symptoms using machine learning. In: Sikander A, Acharjee D, Chanda CK, Mondal PK, Verma P, editors. Energy Systems, Drives and Automations: Proceedings of ESDA 2019. Singapore: Springer; 2020. pp. 561–569.

9. Kohli PS, Arora S. Application of machine learning in disease prediction. In: 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, December 14–15, 2018. pp. 1–4.
10. Patil S, Jaybhaye S, Bokariya S, Jain P, Phapale S, Hande T. Parkinson's disease prediction system in machine learning. ITM Web Conf. 2023; 56: 05002.
11. Kaptoge S, Pennells L, De Bacquer D, Cooney MT, Kavousi M, Stevens G, Riley LM, Savin S, Khan T, Altay S, Amouyel P. World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. Lancet Global Health. 2019; 7 (10): e1332–e1345.
12. Ulianova S. Cardiovascular disease dataset. [Online]. Kaggle.com. 2019. Available at https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset
13. Akturk M. Diabetes dataset. [Online]. Kaggle.com. 2020. Available at https://www.kaggle.com/datasets/mathchi/diabetes-data-set
14. Mayo Clinic. Parkinson's disease – symptoms and causes. [Online]. Mayo Clinic. 2025. Available at https://www.mayoclinic.org/diseases-conditions/parkinsons-disease/symptoms-causes/syc-20376055
15. Ukani V. Parkinson's disease dataset. [Online]. Kaggle.com. 2019. Available at https://www.kaggle.com/datasets/vikasukani/parkinsons-disease-data-set
16. Islam MA, Majumder MZ, Hussein MA. Chronic kidney disease prediction based on machine learning algorithms. J Pathol Informatics. 2023; 14: 100189.
17. University of California, Irvine. UCI machine learning repository. [Online]. Uci.edu. 2015. Available at https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease
18. Jović A, Brkić K, Bogunović N. A review of feature selection methods with applications. In: 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, May 25–29, 2015. pp. 1200–1205.
19. Sarker IH.. Machine learning: algorithms, real-world applications and research directions. SN Computer Sci. 2021; 2 (3): 160.
20. LeCessie S, Van Houwelingen JC. Ridge estimators in logistic regression. J R Stat Soc Ser C (Appl Stat). 1992; 41 (1): 191–201.
21. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011; 12: 2825–2830.
22. Steinbach M, Tan PN. kNN: k-nearest neighbors. In: The Top Ten Algorithms in Data Mining 2009. London, UK: Chapman & Hall/CRC Press; 2009. pp. 165–176.
23. IBM. Support vector machine. [Online]. Ibm.com. 2023. Available at https://www.ibm.com/think/topics/support-vector-machine
24. Podgorelec V, Kokol P, Stiglic B, Rozman I. Decision trees: an overview and their use in medicine. J Med Syst. 2002; 26: 445–463.
25. Mahesh B. Machine learning algorithms – a review. Int J Sci Res. 2020; 9 (1): 381–386.
26. Chiu MH, Yu YR, Liaw HL, Chun-Hao L. The use of facial micro-expression state and tree-forest model for predicting conceptual-conflict based conceptual change. In: Lavonen J, Juuti K, Lampiselkä J, Uitto A, Hahl K, editors. Science Education Research: Engaging Learners for a Sustainable Future (ESERA Eproceeding). Helsinki, Finland: ESERA; 2015.,
27. Han J, Pei J, Tong H. Data Mining: Concepts and Techniques. San Francisco, CA, USA: Morgan Kaufmann; 2022.
28. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011; 12: 2825–2830.
29. Wang W, Chakraborty G, Chakraborty B. Predicting the risk of chronic kidney disease (CKD) using machine learning algorithm. Appl Sci. 2020; 11 (1): 202. Available at https://www.researchgate.net/publication/348025909/figure/fig2/AS:1020217916416002@1620250314481/Simplified-structure-of-XGBoost.ppm
30. Anshul. Guide on AdaBoost algorithm. [Online]. Analytics Vidhya. 2021. Available at https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/

