Journal of Young Pharmacists, 2025; 17(1):226-233.
https://www.jyoungpharm.org

Original Article

Cardiovascular Disease Prediction Using Machine Learning Metrics

Aashish Gnanavelu[1], Champa Venkataramu[2], Ramakrishna Chintakunta[3,*]

[1]Department of Pharmacy Practice, PESU Institute of Pharmacy, PES University, Electronic City Campus, Bengaluru, Karnataka, INDIA.
[2]Department of Computer Applications, SSMRV College, BCU University, Jayanagar, Bengaluru, Karnataka, INDIA.
[3]Department of Pharmaceutical Chemistry, PESU Institute of Pharmacy, PES University, Electronic City Campus, Bengaluru, Karnataka, INDIA.

ABSTRACT

Background: This project aims to develop a Machine-Learning model for heart disease prediction based on clinical and demographic data. Traditional diagnostic methods may not always detect subtle risk factors, hence Machine Learning offers a promising approach to enhance predictive accuracy. **Materials and Methods:** The methodology involves comprehensive pre-processing of the Kaggle Heart Disease dataset, applying algorithms such as Decision Tree, K-Nearest Neighbors, Naive Bayes algorithm, XGBoost, and Random Forest for predictive modelling. **Results:** The XGBoost algorithm outperformed other models with an accuracy of 93% on the test set. Key predictors of heart disease identified through feature importance analysis included age, sex, BMI, genetic, and lifestyle factors. An interactive dashboard was developed to enable users to predict the likelihood of heart disease based on input parameters. **Conclusion:** This project demonstrates the feasibility and effectiveness of Machine Learning techniques in predicting heart disease, enabling timely interventions and personalized treatment strategies. Future research directions include integrating additional data sources and refining models to improve prediction accuracy and robustness in diverse patient populations.

Keywords: Data Analysis, Healthcare, Heart Disease Prediction, Machine Learning Algorithms, Predictive Accuracy.

Correspondence:
Dr. Ramakrishna Chintakunta
Department of Pharmaceutical Chemistry, PESU Institute of Pharmacy, PES University, Electronic City Campus, Bengaluru-560100, Karnataka, INDIA.
Email: ramakrishna@pes.edu
ORCID: 0000-0001-9016-0400

Received: 10-05-2024;
Revised: 12-09-2024;
Accepted: 02-12-2024.

INTRODUCTION

Heart disease is a major global health concern, causing around 17.9 million deaths annually, accounting for 31% of all global deaths. Predicting heart disease is crucial, as it can help mitigate its devastating impact on individuals and healthcare systems. Early detection can improve patient outcomes by preventing complications and reducing mortality rates. Predictive models can also help healthcare providers to optimize resource allocation and personalized treatment strategies, thereby enhancing healthcare delivery efficiency. Traditional risk assessment methods, like the Framingham Risk Score, may not capture all relevant variables or interactions. Machine Learning offers a transformative approach by analysing vast patient data to identify complex patterns and relationships, improving predictive accuracy, and individualizing risk assessment based on a more comprehensive range of factors. Predictive models can inform population-level interventions and policy decisions aimed at reducing the burden of heart disease, identifying high-risk groups, and targeting preventive measures like lifestyle modifications and early screening programs. This could potentially lower healthcare costs and improve overall health outcomes on a broader scale.[1,2]

Heart disease is a major global mortality cause, necessitating effective predictive models for early detection and intervention strategies. Advancements in Machine Learning (ML) have led to the development of accurate models for predicting cardiovascular outcomes based on patient data. Early studies, like the Framingham Heart Study, established traditional risk factors as crucial predictors of cardiovascular risk. Machine Learning techniques have emerged as powerful tools for leveraging complex relationships among predictors to improve prediction accuracy. Decision tree algorithms, such as CART, have segmented patient populations based on risk factors, offering insights into heterogeneous risk profiles. Random forest algorithms have gained popularity for their ability to handle high-dimensional data and mitigate overfitting, improving predictive performance in cardiovascular risk prediction tasks.[3,4]

However, challenges persist in current literature, such as the availability of homogeneous datasets and interpretability concerns. Understanding the clinical relevance of model predictions and translating them into actionable insights for healthcare providers is crucial for integrating ML-based tools

OPEN ACCESS

Manuscript TECHNOMEDIA

DOI: 10.5530/jyp.20251231

Publishing Partner : Manuscript Technomedia. [www.mstechnomedia.com]

into routine clinical practice. This literature review highlights the evolving landscape of heart disease prediction using Machine Learning, highlighting the progress made and persistent challenges that must be addressed. The hypothetical project aims to fill this gap by focusing on comparative evaluations of ML algorithms using diverse datasets and enhancing interpretability for clinical utility.[5,6]

While existing literature on heart disease prediction using Machine Learning has made significant strides, several critical gaps persist. One notable gap is the lack of comprehensive studies that systematically compare and evaluate multiple Machine Learning algorithms using diverse datasets. Many current studies often focus on specific populations or utilize limited datasets, which may compromise the generalizability and applicability of their findings across broader patient demographics. Moreover, there is a need for greater emphasis on the interpretability and clinical relevance of predictive models, alongside their predictive accuracy. Addressing these gaps is crucial for advancing the field towards more robust and clinically applicable predictive tools that can effectively guide personalized healthcare interventions and improve patient outcomes.[7,8]

The process of developing a predictive model for heart disease involves several steps. First, a comprehensive dataset of patient records is collected and pre-processed to ensure its suitability for Machine Learning models. Next, Exploratory Data Analysis (EDA) is conducted to visualize the distribution and relationships between variables like age, blood pressure, and cholesterol levels. Multiple Machine Learning models are developed, such as logistic regression, decision trees, random forests, and XGBoost, to predict heart disease. The performance of these models is evaluated using metrics like accuracy, precision, recall, and F1-score. The models are then compared with traditional risk assessment methods to validate improvements in predictive accuracy and clinical relevance. A user-friendly interface is created to allow healthcare professionals to input patient data and receive predictions. The model is then validated with new data to ensure its generalizability and robustness. Interactive dashboards or user interfaces are created to summarize complex model outputs in a user-friendly manner.[9,10]

The project will use a dataset of anonymized patient records from a cardiovascular clinic or healthcare database, including demographic information, physiological measurements, and clinical history. Machine Learning techniques will be used to develop predictive models for heart disease, including decision trees, K-nearest neighbors, the Naive Bayes algorithm, XGBoost, random forests, and potentially deep learning models. Feature selection and engineering techniques will be applied to identify and prioritize relevant predictors of heart disease risk. The performance of the developed models will be evaluated using standard metrics. The project will assess the practical applicability of the predictive models in clinical settings, considering factors like interpretability and integration into existing healthcare practices. Recommendations for mode deployment and potential benefits in enhancing clinical decision-making will be explored. The project scope includes limitations such as data availability and quality, potential biases, and generalizability of findings to diverse populations and healthcare settings. Ethical considerations regarding data privacy and patient consent will be adhered to throughout the project.[11,12]

MATERIALS AND METHODS

Tools and Technologies

Hardware: Local Machine: Development on a laptop with an Intel i5 processor and 16GB RAM.

Software and Libraries: Programming Language (Python 3.8): The primary programming language used for data analysis and Machine Learning.

Pandas: For data manipulation and analysis.

NumPy: For numerical operations.

Scikit-learn: For Machine Learning algorithms and data preprocessing.

Matplotlib and Seaborn: For data visualization.

SMOTE stands for Synthetic Minority Oversampling Technique, is used primarily for handling imbalanced datasets in Machine Learning.

Jupyter Notebook: For interactive coding and documentation.

Methods

Data collection

This study collected clinical and demographic data from patients during routine medical examinations, including information on heart disease, BMI, smoking, alcohol consumption, stroke, physical and mental health, and physical activity, adhering to ethical guidelines and patient confidentiality protocols.[13]

Data pre-processing

The data processing phase involves several steps to ensure data quality and consistency. These include handling missing values, identifying unique values, removing duplicate values, encoding categorical variables, distinguishing between categorical and numerical variables, and detecting outliers.[14] These steps are crucial to ensure data integrity and avoid bias in model training.

1. Missing values are identified using techniques like imputation or deletion based on the percentage of missing data and variable nature.
2. Unique values are identified by exploring each variable, particularly categorical variables, to understand their range and distribution.

3. Duplicate values are removed to ensure data integrity and avoid bias in model training.

4. Categorical variables, such as sex and smoking status, are encoded using techniques like LabelEncoder() and fit.transform().

5. Outliers are detected using statistical methods and domain knowledge, and appropriate actions are taken to mitigate their impact on model performance.

Exploratory Data Analysis (EDA)

It involves statistical analysis, visualization, and computation of summary statistics, correlations, and target variable analysis to understand data distributions and relationships.[15]

Feature Engineering

The feature engineering process involved calculating Pearson correlation coefficients to identify highly correlated features with the target variable, prioritizing these features accordingly.

Model selection

The selection of Machine Learning models for classification tasks in healthcare analytics was based on their suitability for handling dataset size and complexity. Decision trees, random forests, K-Nearest Neighbors (KNN), XGBoost, and Naive Bayes were evaluated for their ability to capture complex relationships within data. Performance metrics such as accuracy, precision, recall, f1-score, support, computational efficiency, and interpretability were also considered.[16,17]

Training and Testing

The dataset was divided into training and testing sets for robust model evaluation. The 70% training set was used for unseen

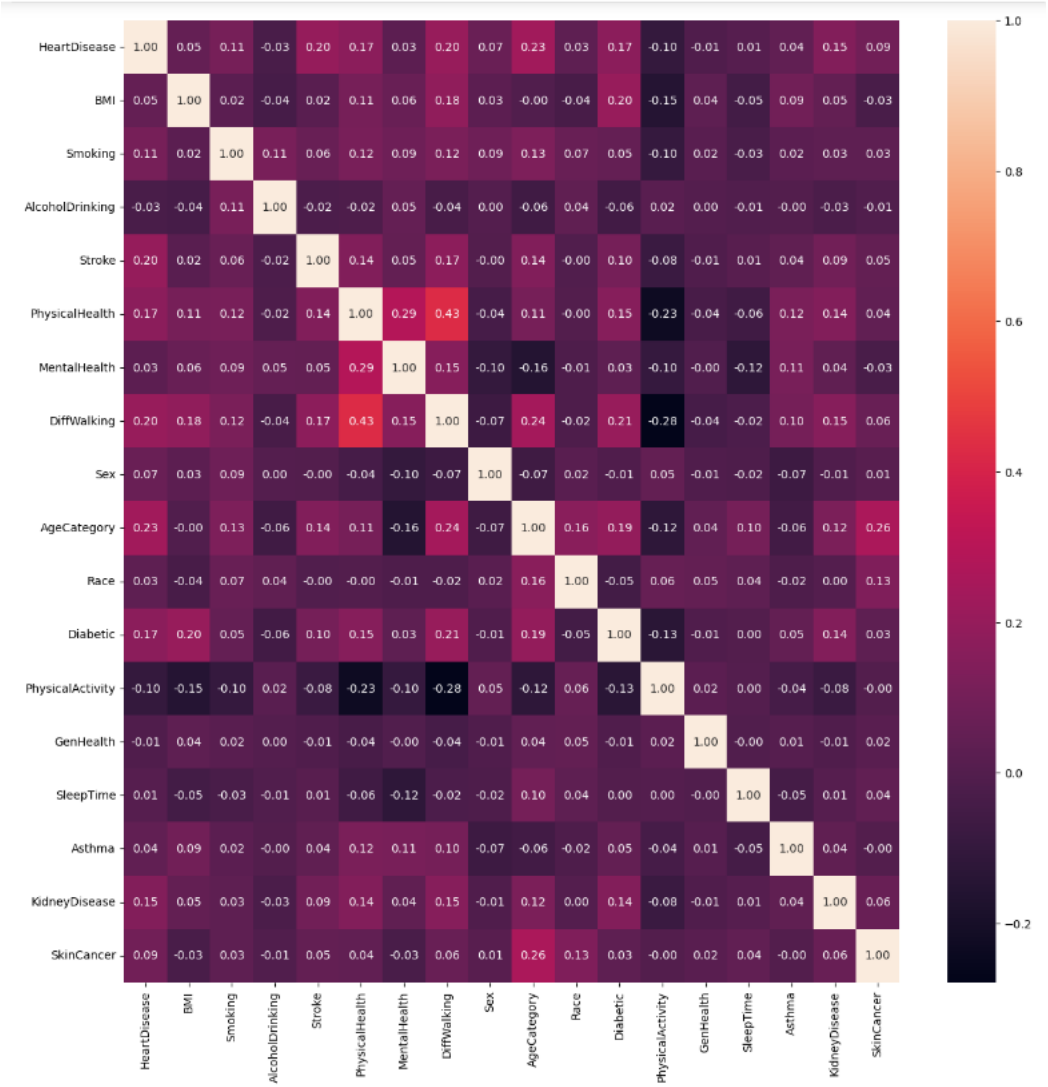

Figure 1: Correlation Analysis.

data, while 30% was used for testing. Each model was trained using optimal hyperparameters with multiple iterations to avoid overfitting. The trained models were evaluated using various metrics, including accuracy, precision, recall, F1-score, and support. The ensemble model was also tested to compare its performance against individual models.[18]

Model Evaluation

Evaluate the model's performance using appropriate metrics for classification tasks, such as;

Accuracy: Overall correctness of predictions.

Precision: Proportion of true positive predictions among all positive predictions.

Recall (sensitivity): Proportion of true positive predictions among all actual positives.

F1-score: Harmonic mean of precision and recall, providing a balance between them.

Model Validation and Selection

The process involves evaluating the best-performing models on a validation set and comparing their performance metrics to select the final model.[19]

User Interface Design (Dashboard Components)

Prediction Interface: Allow users to input health parameters (e.g., BMI, smoking, age, etc.) to receive predictions on heart disease presence.

Filters and Controls: Enable users to filter data based on demographic variables(sex, age category, race) or health parameters (physical activity, sleep time).

RESULTS

Correlation Analysis

Pearson correlation coefficients were calculated to identify features highly correlated with the target variable.[20] Features with high correlation were prioritized. Mentioned in Figure 1.

The performance of Machine Learning techniques is evaluated using four parameters: recall, F1-measure, accuracy, and precision. A confusion matrix is used to measure the potential of these parameters, representing the number of subjects correctly classified as "positive"(heart disease presence), "negative"(absence/healthy heart disease), "negative"(heart disease absence), "negative"(heart disease), and "positive"(heart disease absence). The number of subjects incorrectly classified as "negative" is represented by Fn and Fp.

Accuracy: The proportion of correctly classified instances out of the total instances.

Formula: Accuracy=True Positives+True Negatives/Total Instances.

Interpretation: High accuracy indicates better overall performance but does not distinguish between types of errors.

Precision: The proportion of true positive predictions out of all positive predictions.

Formula: Precision=True Positives/True Positives+False Positives.

Interpretation: High precision indicates a low false positive rate.

Recall: The proportion of true positive predictions out of all actual positive instances.

Formula: Recall=True Positives/True Positives+False Negatives.

Interpretation: High recall indicates a low false negative rate.

F1 Score: The harmonic mean of precision and recall, balancing the two metrics.

Formula: F1 Score=2 × Precision × Recall/Precision+Recall.

Interpretation: A high F1 score indicates a good balance between precision and recall.

Different Machine Learning heart disease prediction parameters given in Table 1 and different Machine Learning techniques of precision and accuracy given in Figure 2 and heart disease prediction in terms of age category given in Figure 3.

After evaluating multiple Machine Learning models for predicting heart disease, our study found that the XG Boost model achieved the highest accuracy of 93%, surpassing our initial objective of 71% accuracy. This model also demonstrated strong performance in precision (97%) and recall (88%), indicating its ability to effectively identify both positive and negative cases of heart disease. Importantly, age and cholesterol levels emerged as the most influential features in predicting heart disease risk, aligning

Table 1: Different Machine Learning model-based heart disease detection is evaluated using different parameters.

Model	Accuracy	Precision	Recall	F1-score
Decision Tree	0.81	0.79	0.84	0.81
KNN	0.87	0.81	0.97	0.88
Naive Bayes	0.71	0.75	0.62	0.68
XGBoost	0.93	0.97	0.88	0.93
Random Forest	0.82	0.81	0.85	0.83

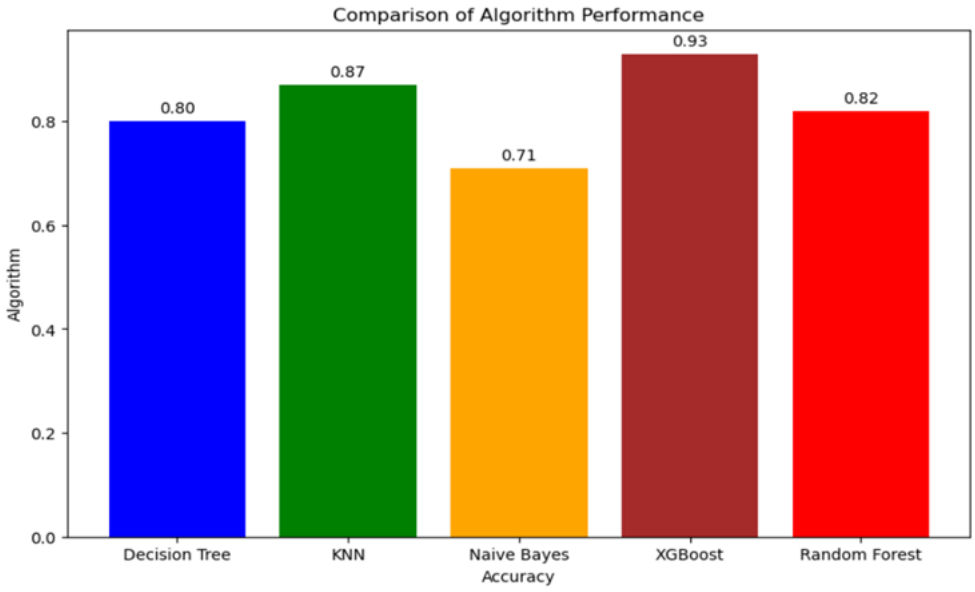

Figure 2: Different Machine Learning techniques used to predict heart disease in terms of accuracy.

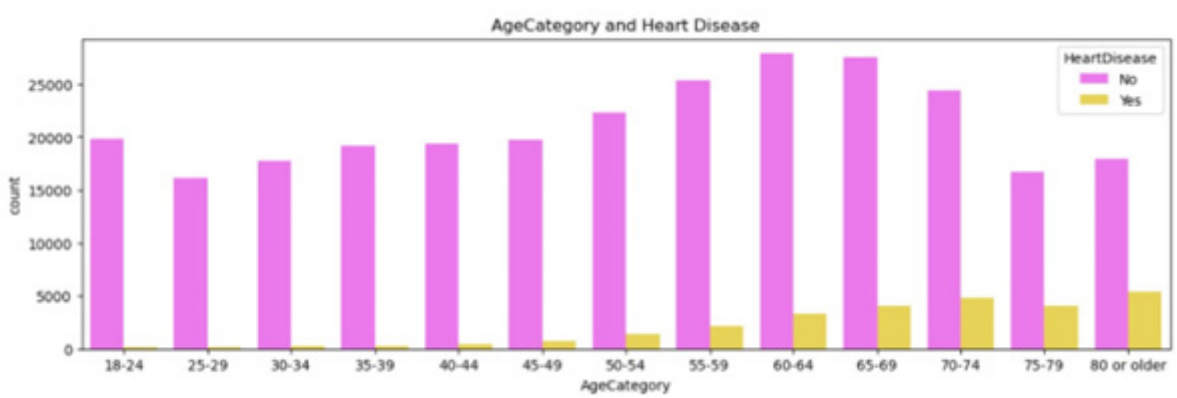

Figure 3: Prediction of heart disease in terms of age category parameter.

with established medical literature. While our study focused on a specific demographic, future research could explore the generalizability of these findings across broader populations and consider incorporating genetic data for enhanced predictive accuracy. Development of interactive dash board is given in Figure 4.

DISCUSSION

The primary objectives include exploring various Machine Learning algorithms, pre-processing the dataset to handle missing values and outliers, model selection, and evaluating the model's predictive accuracy using appropriate metrics on readily available patient information.

The study compared various models, including Decision Trees, Random Forests, Naive Bayes, K-Nearest Neighbors, and Gradient Boosting, to determine their performance in heart disease prediction tasks. The results showed that these models had lower performance metrics, indicating their limitations in capturing the complexity of heart disease prediction.

The XG Boost and KNN models demonstrated robustness in predictions, demonstrating high accuracy and precision-recall values, indicating their ability to accurately differentiate between individuals with and without heart disease.

Machine Learning models can assist healthcare professionals in early heart disease detection and risk assessment, enabling clinicians to prioritize interventions and treatments for at-risk patients more effectively. The models' reliability and applicability in real-world healthcare scenarios could be enhanced through further improvement through feature engineering, advanced ensemble methods, optimization of model hyperparameters, and further validation on diverse datasets and clinical settings.

The use of Machine Learning (ML) in heart disease prediction has significant implications for healthcare and the industry. ML models, such as gradient boosting and random forest, can facilitate

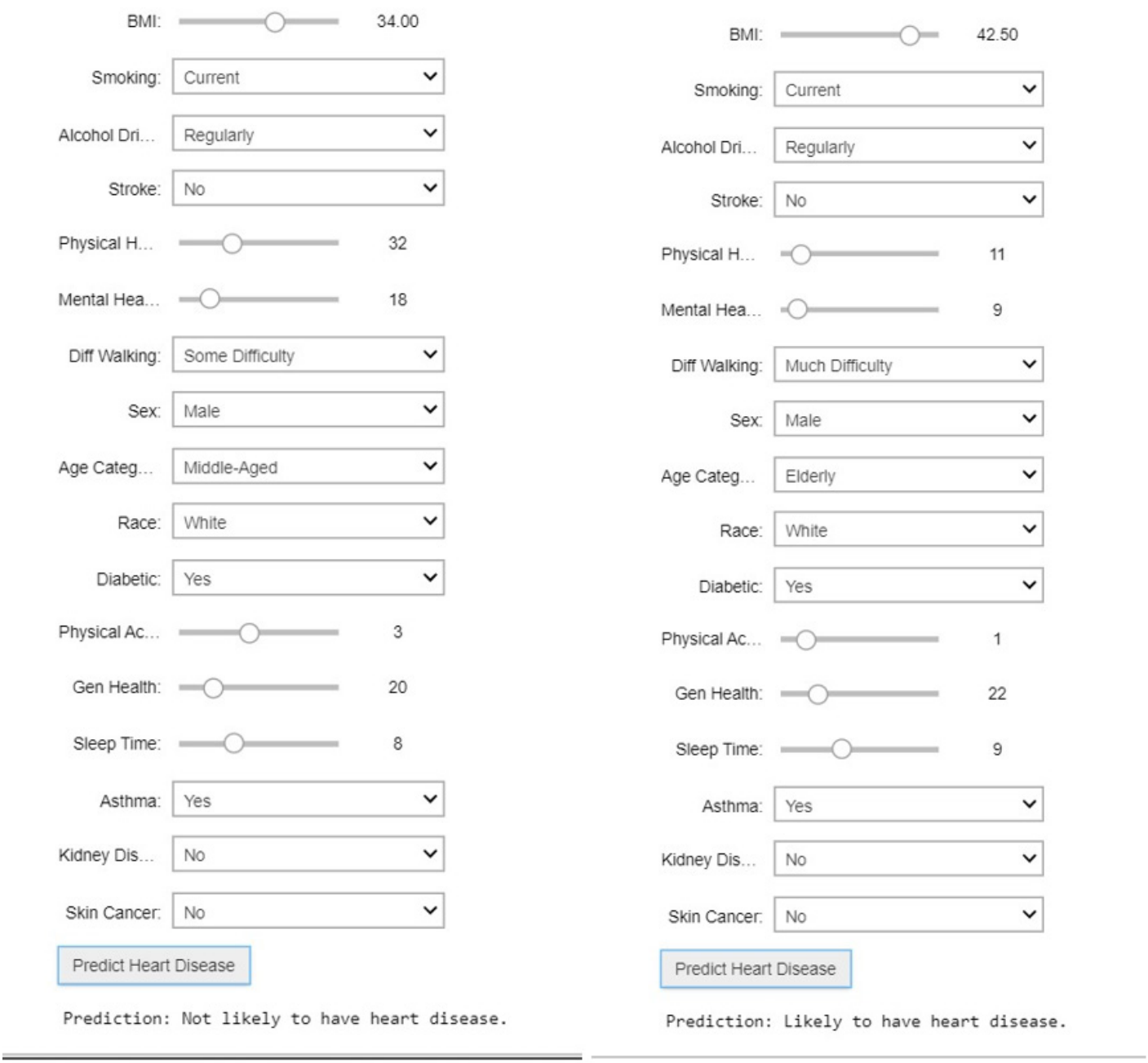

Figure 4: Developing an interactive dashboard for heart disease prediction involves integrating Machine Learning models with user-friendly visualizations for informed health insights parameters.

early detection of heart disease, allowing healthcare providers to intervene proactively. This can reduce morbidity, mortality, and healthcare costs associated with advanced cardiovascular conditions.

ML-driven predictive models also enable early identification of at-risk individuals, potentially reducing disease burden and associated healthcare costs. These models support public health initiatives aimed at improving population health outcomes.

Personalized medicine is another area where ML models can be used. They can integrate diverse patient data to generate personalized risk assessments, enabling clinicians to prioritize interventions and treatments based on individualized risk profiles. These models also serve as powerful tools for clinical decision support, providing data-driven insights and predictive analytics. Accurate prediction models help allocate healthcare resources more efficiently, focusing on high-risk populations and optimizing resource utilization. These findings contribute to

ongoing research efforts in cardiovascular medicine, guiding the development of novel diagnostic tools, therapeutic strategies, and preventive measures.

Advanced feature engineering techniques can capture complex interactions and nonlinear relationships between predictors, using domain-specific knowledge and domain adaptation techniques to refine feature selection and extraction processes. Enhanced model interpretability can be improved by exploring techniques like SHAP values and LIME to provide insights into model predictions and enhance trust among healthcare professionals. Longitudinal data analysis can be used to capture disease progression and temporal changes in risk factors, enabling dynamic predictive models that can adapt over time.

Clinical validation and adoption should be conducted through prospective studies and collaboration with healthcare providers to integrate ML-driven decision-support tools into routine clinical workflows. Ethical considerations related to data privacy, patient consent, and algorithmic fairness should be addressed, and bias detection and mitigation strategies should be implemented to ensure equitable and unbiased predictions across different demographic groups.

Collaborative research initiatives should be promoted to address complex challenges in cardiovascular disease prediction and management. Personalized risk assessment can be developed by developing models that can dynamically adapt to individual patient profiles over time, incorporating genetic, epigenetic and omics data to enhance predictive models and uncover novel biomarkers for cardiovascular risk assessment.

Advanced model architectures, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and attention mechanisms, can be explored for feature extraction and prediction in heart disease. Interpretable AI in healthcare can be developed and refined, providing transparent insights into decision-making processes. Validation and real-world deployment should involve large-scale prospective studies and clinical trials to validate the effectiveness and clinical utility of ML models in real-world healthcare settings.

However, challenges such as data quality, model interpretability, and regulatory compliance remain crucial for the widespread adoption of ML in healthcare. Opportunities lie in further refining ML models, integrating them into clinical workflows, and fostering interdisciplinary collaborations. Ethical considerations such as patient privacy, data security, and transparency are also essential for the trust and acceptance of ML technologies in healthcare.

CONCLUSION

The XG Boost model consistently outperformed other models in terms of accuracy, precision, recall, and F1 score, demonstrating its ability to accurately predict individuals with and without heart disease while maintaining high precision and recall rates. It outperformed decision trees, random forests, naive bayes, and K-nearest neighbors across all evaluated metrics.

ML has transformative potential in healthcare by enhancing clinical decision-making, supporting personalized medicine initiatives, and improving overall healthcare outcomes. Future research in heart disease prediction using Machine Learning (ML) should focus on integrating diverse data sources, such as genetic information, lifestyle factors, and environmental variables, to enhance predictive models. This can be achieved by combining structured clinical data with unstructured data like medical imaging and free-text clinical notes.

ACKNOWLEDGEMENT

We are thankful to the professors, students, and management of SSMRV College and PES University Bengaluru for the completion of this article.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ABBREVIATIONS

ML: Machine Learning; **EDA:** Exploratory Data Analysis; **KNN:** K Nearest Neighbours; **XGBoost:** Extreme Gradient Boosting.

REFERENCES

1. Salhi DE, Tari A, Kechadi T. Using machine learning for heart disease prediction; 2021. doi: 10.1007/978-3-030-69418-0_7.
2. Ferjani M. Disease prediction using. Mach Learn. 2020. doi: 10.13140/RG.2.2.18279.47521.
3. Bhatt CM, Patel P, Ghetia T, Mazzeo PL. Effective heart disease prediction using machine learning techniques. Algorithms. 2023;16(2). doi: 10.3390/a16020088.
4. Muhammad Y, Tahir M, Hayat M, Chong KT. Early and accurate detection and diagnosis of heart disease using intelligent computational model. Sci Rep. 2020;10(1):19747. doi: 10.1038/s41598-020-76635-9, PMID 33184369.
5. Ahsan MM, Luna SA, Siddique Z. Machine-learning-based disease diagnosis: A comprehensive Review [review]. Healthcare (Basel). 2022;10(3):541. doi: 10.3390/healthcare10030541, PMID 35327018, PMCID PMC8950225.
6. Logabiraman G, Ganesh D, Kumar MS, Kumar AV, Bhardwaj N. Heart disease prediction using machine learning algorithms. MATEC Web Conf. 2024;392. doi: 10.1051/matecconf/202439201122.
7. Ahmad AA, Polat H. Prediction of heart disease based on machine learning using jellyfish optimization algorithm. Diagnostics (Basel). 2023;13(14):2392. doi: 10.3390/diagnostics13142392, PMID 37510136, PMCID PMC10378171.
8. Nagavelli U, Samanta D, Chakraborty P. Machine learning technology-based heart disease detection models. J Healthc Eng. 2022; 2022:7351061. doi: 10.1155/2022/7351061, PMID 35265303, PMCID PMC8898839.
9. Ye Z, An S, Gao Y, Xie E, Zhao X, Guo Z et al. The prediction of in-hospital mortality in chronic kidney disease patients with coronary artery disease using Machine Learning models. Eur J Med Res. 2023;28(1):33. doi: 10.1186/s40001-023-00995-x, PMID 36653875.
10. Krittanawong C, Virk HU, Bangalore S, Wang Z, Johnson KW, Pinotti R, *et al.* Machine Learning prediction in cardiovascular diseases: a meta-analysis. Sci Rep. 2020;10(1):16057. doi: 10.1038/s41598-020-72685-1, PMID 32994452.
11. Available from: https://www.who.int/news-room/fact-sheets/detail/Cardiovascular-Diseases-(cvds).
12. Maini E, Venkateswarlu B, Maini B, Marwaha D. Machine Learning-based heart disease prediction system for Indian population: an exploratory study done in South India. Med J Armed Forces India. 2021;77(3):302-11. doi: 10.1016/j.mjafi.2020.10.013, PMID 34305284.
13. Chandrasekhar N, Peddakrishna S. Enhancing heart disease prediction accuracy through machine learning techniques and optimization. Processes. 2023;11(4):1210. doi: 10.3390/pr11041210.

14. Rajdhan A, Agarwal A, Sai M, Ravi D, Ghuli P. Heart disease prediction using machine learning. Int J Eng Res Technol. 2020;9(4):659-62.
15. Bhatt CM, Patel P, Ghetia T, Mazzeo PL. Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms. 2023;16(2):88. doi: 10.3390/a16020088.
16. Ramalingam VV, Dandapath A, Karthik Raja M. Heart disease prediction using Machine Learning techniques: A survey. Int J Eng Technol. 2018;7(3):684. doi: 10.14419/ijet.v7i2.8.10557.
17. Jindal H, Agrawal S, Khera R, Jain R, Nagrath P. Heart disease prediction using machine learning algorithms. IOP Conf S Mater Sci Eng. 2021;1022(1):012072. doi: 10.1088/1757-899X/1022/1/012072.
18. Available from: https://news.harvard.edu/gazette/story/2001/04/drinkers-less-likely-to-die-from-heart-attacks/.
19. Kaggle. Dataset. Available from: https://www.kaggle.com/learn.
20. Raschka S, Mirjalili V. Machine Learning and Deep Learning with Python, scikit learn, and TensorFlow 2. 3rd ed ed. Birmingham and Mumbai: Packt Publishing; 2020.

Cite this article: Aashish G, Champa V, Chintakunta R. Cardiovascular Disease Prediction Using Machine Learning Metrics. J Young Pharm. 2025;17(1):226-33.