IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Heart Disease Prediction Using Machine Learning Techniques

Dr. Santosh Kumar Singh1, Asst. Prof. Rimsy Dua2

Mr. Saad Mulla3, Mr. Niraj Jadhav4

1.Head Of Dept., 2.Assistant Professor, 3,4 PG Student

Department of IT, Thakur College of Science and Commerce, Thakur Village

Kandivali (East), Mumbai, Maharashtra, India

Abstract : Heart disease remains one of the leading causes of mortality worldwide. Early diagnosis and effective risk assessment are crucial for preventing severe complications. Machine learning (ML) techniques have emerged as powerful tools in medical diagnostics, providing automated and efficient predictions based on patient data. This study explores the application of various ML models, including Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Random Forest, XGBoost, and Neural Networks, to predict heart disease. The models are trained on a dataset of 303 patient records with 14 features, and their performances are evaluated. The results indicate that the Random Forest algorithm achieves the highest accuracy of 95.08%, outperforming other models. The study emphasizes the significance of ML in medical decision-making and highlights the potential of ensemble learning methods in improving predictive accuracy.

IndexTerms - Heart Disease, Machine Learning, Prediction, Classification, Medical Diagnosis, Artificial Intelligence, Healthcare Analytics

INTRODUCTION

Heart disease is a major public health concern worldwide, accounting for millions of deaths each year. Traditional diagnostic approaches, such as electrocardiograms, angiography, and echocardiography, require specialized equipment and expertise, making them time-consuming and expensive. In contrast, machine learning provides a data-driven approach that can improve early diagnosis, reduce diagnostic errors, and assist healthcare professionals in making informed decisions. This paper aims to explore multiple ML techniques and their effectiveness in predicting heart disease.

The integration of ML into healthcare analytics is revolutionizing the way diseases are diagnosed. Researchers have demonstrated that ML algorithms can process vast amounts of medical data, detect patterns, and predict diseases with high accuracy. This not only improves patient care but also reduces the burden on healthcare systems. The goal of this study is to provide a comprehensive analysis of different ML algorithms, comparing their performances in heart disease prediction and highlighting the best-performing models.

2. Objectives The primary objectives of this study are:

1. To analyze the performance of different machine learning models in predicting heart disease.
2. To identify the most relevant features contributing to heart disease prediction.
3. To compare various classification models based on accuracy, precision, recall, and F1-score.

4. To determine the effectiveness of ensemble learning techniques in improving predictive accuracy.
5. To propose a robust ML model that can assist healthcare professionals in diagnosing heart disease efficiently.

3. Methodology

The methodology of this study consists of several key phases, including dataset selection, preprocessing, feature selection, model implementation, training, and evaluation. Each phase plays a crucial role in ensuring the reliability and accuracy of the predictive models.

3.1 Dataset The study utilizes the UCI Heart Disease dataset, which consists of 303 patient records with 14 attributes. The dataset includes key health indicators such as age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression, the slope of the peak exercise ST segment, number of major vessels, and thalassemia type. The target variable indicates the presence (1) or absence (0) of heart disease.

3.2 Data Preprocessing To ensure data quality and enhance model performance, several preprocessing steps were performed:

• Handling Missing Values: The dataset was checked for missing values, and none were found.
• Normalization and Scaling: Continuous numerical features were normalized using Min-Max Scaling to standardize the data range.
• Encoding Categorical Variables: Categorical attributes such as chest pain type and thalassemia were converted into numerical representations using one-hot encoding.
• Correlation Analysis: The correlation between independent variables and the target variable was analyzed to identify the most significant predictors.
• Data Splitting: The dataset was divided into an 80% training set and a 20% test set to evaluate model generalization.
• Feature Selection: A feature selection technique, such as Recursive Feature Elimination (RFE), was applied to reduce dimensionality and improve model performance.

3.3 Machine Learning Models This study implements and evaluates the following machine learning models:

• Logistic Regression (LR): A binary classification algorithm that predicts probabilities based on a logistic function.
• Naïve Bayes (NB): A probabilistic model that assumes independence between features and applies Bayes' theorem for classification.
• Support Vector Machine (SVM): A model that identifies an optimal hyperplane to maximize the margin between two classes.
• K-Nearest Neighbors (KNN): A distance-based model that classifies instances based on the majority vote of their nearest neighbors.
• Decision Tree (DT): A tree-structured model that recursively splits the dataset into branches based on feature values.
• Random Forest (RF): An ensemble of multiple decision trees that improves classification performance by reducing overfitting.
• XGBoost: A gradient boosting algorithm that optimizes prediction performance by sequentially correcting previous errors.
• Neural Networks (NN): A deep learning-based model that utilizes multiple layers of interconnected neurons to learn complex patterns.

3.4 Model Training and Evaluation Each model was trained using the training dataset and evaluated on the test dataset. The following performance metrics were used:

• Accuracy: Measures the percentage of correctly classified instances.
• Precision: Indicates the proportion of true positive cases among all predicted positive cases.
• Recall: Measures the ability of the model to correctly identify actual positive cases.
• F1-Score: A harmonic mean of precision and recall to balance both measures.
• ROC-AUC Score: Evaluates the ability of the model to distinguish between classes.

3.5 Hyperparameter Tuning To improve model performance, hyperparameter tuning was conducted using Grid Search and Random Search techniques. Key parameters such as learning rate, number of trees (for ensemble models), and kernel functions (for SVM) were optimized.

3.6 Cross-Validation To ensure model robustness, k-fold cross-validation (k=10) was performed. This technique reduces variance by training and testing the model on different subsets of the data multiple times.
year for KSE-100 index.

4. Results and Discussion

4.1 Accuracy Comparison

Model	Accuracy (%)
Logistic Regression	85.25
Naïve Bayes	85.25
SVM	81.97
KNN	67.21
Decision Tree	81.97
Random Forest	95.08
XGBoost	85.25
Neural Network	80.33

The results show that Random Forest achieves the highest accuracy of 95.08%, making it the most reliable model for heart disease prediction. Ensemble methods like Random Forest and XGBoost demonstrate superior performance compared to individual classifiers due to their ability to reduce overfitting and improve generalization.

5. Conclusion and Future Work This research highlights the efficacy of machine learning models in predicting heart disease, with Random Forest demonstrating the highest accuracy. The study emphasizes the importance of feature selection, data preprocessing, and ensemble learning in enhancing predictive performance. Future work should explore:

- Larger datasets with diverse demographics for better generalization.
- Integration of deep learning models like CNNs and RNNs for improved feature extraction.
- Development of real-time predictive systems for clinical applications.
- Incorporating explainable AI (XAI) techniques to improve model interpretability.
- Evaluating the impact of feature engineering techniques on predictive accuracy.

Literature review:

Shanmugasundaram G. et al. conducted a comprehensive investigation into heart disease prediction techniques, emphasizing the role of data mining and machine learning in improving diagnostic accuracy. Various classification algorithms, including Naïve Bayes, Decision Trees, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and hybrid models, have been explored for predicting heart disease based on both common and medical-oriented factors. Common factors such as age, gender, smoking, alcohol consumption, and obesity, along with medical parameters like blood pressure, cholesterol levels, ECG results, and thallium scans, have been used as key predictors. Studies reviewed in this research indicate that Naïve Bayes, Decision Trees, and Neural Networks are widely utilized, with some models achieving over 90% accuracy. However, challenges remain, including the exclusion of critical attributes, insufficient error-handling techniques, and dataset limitations. Some researchers have applied dimensionality reduction for efficiency, but this has sometimes compromised prediction accuracy. The study suggests that future research should focus on incorporating all influencing factors and leveraging advanced machine learning techniques such as deep learning and ensemble methods to enhance predictive precision and reliability in heart disease diagnosis.[1]

Jyoti Soni et al. conducted a comprehensive review on heart disease prediction using predictive data mining techniques, highlighting the growing importance of machine learning in medical diagnosis. The study explores various classification algorithms, including Decision Trees, Naïve Bayes, K-Nearest Neighbors (KNN), Neural Networks, and Classification via Clustering, to evaluate their effectiveness in heart disease prediction. Their findings indicate that Decision Trees outperform other techniques in accuracy and interpretability, while Bayesian classification exhibits comparable performance in some cases. Additionally, the application of

genetic algorithms enhances the accuracy of Decision Trees and Bayesian classifiers by optimizing feature selection. The study also discusses the limitations of certain predictive models, such as KNN and Neural Networks, which struggle with noisy and high-dimensional data. Furthermore, association rule mining and rough set theory are explored to refine rule-based classification and improve interpretability. Despite significant advancements, challenges remain, including the need for better data preprocessing, integration of multiple classifiers, and validation on real-world datasets. The research underscores the necessity for developing an automated and efficient heart disease prediction system by leveraging a combination of machine learning models, genetic algorithms, and feature selection techniques to enhance diagnostic accuracy and clinical decision-making.[2]

A. J. Singh and Mukesh Kumar conducted a comparative analysis on software effort estimation using machine learning techniques, emphasizing the need for accurate predictions in software development. The study explores various machine learning algorithms, including Linear Regression (LR), Multilayer Perceptron (MLP), and Random Forest (RF), evaluating their effectiveness using the WEKA toolkit. Their findings indicate that Linear Regression outperforms MLP and RF in terms of accuracy, as measured by correlation coefficients and error metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The study also reviews traditional estimation methods, such as expert judgment and analogy-based techniques, highlighting their limitations compared to machine learning approaches. Additionally, feature selection techniques, including Information Gain and Correlation-based Feature Selection, are discussed to optimize predictive performance. The research further incorporates genetic algorithms to improve model accuracy by selecting relevant attributes. Despite advancements, challenges remain in handling high-dimensional data and integrating hybrid models for better predictions. The authors emphasize that future research should explore deep learning techniques, ensemble methods, and real-world datasets to refine effort estimation models, ultimately improving software project planning and resource allocation.[3]

Harshit Jindal et al. (2021) present a comprehensive study on heart disease prediction using machine learning algorithms, focusing on logistic regression, KNN, and random forest classifiers. The research aims to enhance the accuracy of predicting cardiovascular diseases by leveraging patient medical history, including attributes like age, chest pain, and blood pressure. The authors utilize a dataset from the UCI repository, comprising 304 patient records, and achieve an accuracy of 87.5%, with KNN outperforming other algorithms at 88.52%. The study highlights the importance of data preprocessing and the use of multiple classifiers to improve prediction accuracy. The proposed model not only aids in early diagnosis but also reduces medical costs by minimizing the need for extensive tests. The results demonstrate that logistic regression and KNN are more effective than random forest in predicting heart disease. The research contributes significantly to the field by providing a cost-efficient and accurate prediction system, which can assist healthcare professionals in diagnosing heart diseases more effectively. However, the study could benefit from exploring larger datasets and incorporating more advanced machine learning techniques to further enhance accuracy and robustness. Overall, the paper offers valuable insights into the application of machine learning in healthcare, particularly in the early detection of cardiovascular diseases.[4]

Purushottam et al. conducted a study on heart disease prediction, emphasizing the importance of automated systems in medical diagnosis to improve accuracy and reduce costs. The research explores various data mining techniques, including the C4.5 decision tree algorithm, Support Vector Machines (SVM), Bayesian classifiers, and Classification Multiple Association Rules (CMAR), to classify heart disease risk levels. The study highlights the use of real-world datasets, such as the Cleveland Heart Disease Database, and evaluates different models using metrics like classification accuracy and rule-based decision-making. The proposed system prioritizes generated rules into different categories—original, pruned, and classified rules—to enhance interpretability. The findings indicate that the proposed heart disease prediction system, implemented using KEEL and WEKA tools, achieves an accuracy of 86.3% in testing and 87.3% in training, outperforming other classifiers. The study emphasizes the need for efficient data preprocessing, integration of multiple classifiers, and feature selection techniques to optimize predictive performance. Future research should focus on refining automated decision support systems using advanced machine learning and deep learning models, incorporating real-time data for enhanced prediction and clinical decision-making, ultimately improving early diagnosis and patient outcomes in healthcare.[5]

M. Marimuthu et al. (2018) provide a comprehensive review of heart disease prediction using machine learning and data analytics approaches. The paper highlights the growing importance of data mining and machine learning techniques in predicting heart diseases, given the increasing complexity of healthcare data. The authors discuss various algorithms such as Artificial Neural Networks (ANN), Decision Trees, K-Nearest Neighbors (KNN), Naive Bayes, and Support Vector Machines (SVM), which are commonly used for heart

disease prediction. The review emphasizes the role of these algorithms in analyzing patient data, including attributes like age, blood pressure, cholesterol levels, and chest pain, to predict the likelihood of heart disease. The paper also summarizes the accuracy of different techniques, noting that SVM and ANN often yield higher accuracy rates compared to other methods. The authors conclude that while significant progress has been made in heart disease prediction, there is still a need for more advanced and hybrid models to improve accuracy and scalability. Future work could explore feature selection methods, multiple classifier voting techniques, and advanced clustering algorithms to enhance prediction performance. Overall, the paper underscores the potential of machine learning and data analytics in revolutionizing healthcare by enabling early and accurate diagnosis of heart diseases, thereby improving patient outcomes and reducing healthcare costs.[6]

V.V. Ramalingam et al. (2018) present a comprehensive survey on heart disease prediction using machine learning techniques, emphasizing the growing importance of accurate and reliable systems for diagnosing cardiovascular diseases (CVDs). The authors highlight that CVDs are a leading cause of global mortality, with significant economic and health impacts, particularly in countries like India. The paper reviews various machine learning algorithms, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes, Decision Trees (DT), Random Forest (RF), and ensemble models, which have been widely used for heart disease prediction. The survey discusses the performance of these algorithms, noting that SVM and Random Forest often achieve high accuracy, with Random Forest reaching up to 97% in some datasets. The authors also explore dimensionality reduction techniques, such as Principal Component Analysis (PCA) and feature selection, which are crucial for handling high-dimensional data and avoiding overfitting. The paper concludes that while machine learning models have shown promising results in predicting heart diseases, there is still a need for further research to improve the handling of complex datasets and to develop more robust ensemble models. The authors suggest that combining multiple algorithms could enhance prediction accuracy and reliability, offering significant potential for advancing healthcare diagnostics and reducing the global burden of heart diseases.[7]

V. Krishnaiah et al. conducted a study on heart disease prediction, emphasizing the role of data mining techniques and intelligent fuzzy approaches in improving diagnostic accuracy. The research explores various classification methods, including Decision Trees, Neural Networks, Naïve Bayes, Support Vector Machines (SVM), and Genetic Algorithms, evaluating their effectiveness in heart disease prediction. The study highlights the use of fuzzy logic to enhance prediction models by handling uncertainties in medical data. It discusses the Intelligent Heart Disease Prediction System (IHDPS), which utilizes historical heart disease datasets to extract hidden patterns and improve decision-making for healthcare professionals. The research also analyzes different data mining techniques, such as clustering, association rule mining, and regression, to identify key risk factors influencing heart disease. The findings suggest that integrating fuzzy logic with machine learning models improves classification accuracy by capturing imprecise and vague medical attributes. The study further emphasizes the importance of optimizing feature selection and developing hybrid models to enhance predictive performance. Future research should focus on incorporating deep learning techniques and real-time data processing to refine heart disease prediction systems, ultimately aiding clinicians in early diagnosis and reducing mortality rates associated with cardiovascular diseases.[8]

Jaymin Patel et al. conducted a study on heart disease prediction using machine learning and data mining techniques, focusing on the effectiveness of Decision Tree algorithms in diagnosing heart disease. The study compares J48, Logistic Model Tree (LMT), and Random Forest algorithms using the Cleveland Heart Disease Dataset from the UCI repository, which contains 303 instances and 76 attributes. The research aims to extract hidden patterns from medical data to enhance disease prediction and reduce unnecessary diagnostic tests. The findings indicate that J48 with reduced-error pruning achieves the highest accuracy and sensitivity, making it the most effective among the tested classifiers. The study also highlights the importance of feature selection, emphasizing that reducing irrelevant attributes improves prediction accuracy. The results show that LMT achieves better specificity, while Random Forest provides stable classification but requires more computational resources. The research concludes that decision trees, particularly J48 with pruning, are highly effective for heart disease prediction, and future studies should explore hybrid models, ensemble techniques, and deep learning approaches to further improve diagnostic precision and scalability in medical applications.[9]

Chaitrali S. Dangare and Sulabha S. Apte (2012) present an improved study on heart disease prediction using data mining classification techniques, focusing on enhancing the accuracy of diagnosis by incorporating additional attributes. The authors highlight the challenges in the healthcare industry, where vast amounts of

data remain unanalyzed, leading to missed opportunities for effective decision-making. The paper proposes a heart disease prediction system that uses 15 input attributes, including two newly added attributes—obesity and smoking—to improve prediction accuracy. The system employs three data mining techniques: Neural Networks, Decision Trees, and Naive Bayes, with Neural Networks achieving the highest accuracy of 100%. The study utilizes the Cleveland and Statlog heart disease datasets, consisting of 573 records, and employs the Weka 3.6.6 tool for data preprocessing and analysis. The results demonstrate that Neural Networks outperform Decision Trees and Naive Bayes, with accuracies of 100%, 99.62%, and 90.74%, respectively. The authors conclude that adding more relevant attributes, such as obesity and smoking, significantly enhances the prediction accuracy of heart disease. They also suggest future work to explore other data mining techniques, such as clustering and association rules, and to incorporate text mining for analyzing unstructured healthcare data. This research contributes to the development of more accurate and reliable heart disease prediction systems, which can aid in early diagnosis and effective treatment, ultimately improving patient outcomes and reducing healthcare costs.[10]

References

1. [1] G. Shanmugasundaram, V. M. Selvam, R. Saravanan and S. Balaji, "An Investigation of Heart Disease Prediction Techniques," 2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA), Pondicherry, India, 2018,
2. [2] Soni J, Ansari U, Sharma D, Soni S. Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications. 2011 Mar
3. [3] Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning techniques. SN Computer Science. 2020 Nov;
4. [4] Jindal H, Agrawal S, Khera R, Jain R, Nagrath P. Heart disease prediction using machine learning algorithms. InIOP conference series: materials science and engineering 2021
5. [5] Saxena K, Sharma R. Efficient heart disease prediction system. Procedia Computer Science. 2016 Jan
6. [6] Marimuthu M, Abinaya M, Hariesh KS, Madhankumar K, Pavithra V. A review on heart disease prediction using machine learning and data analytics approach. International Journal of Computer Applications. 2018 Sep
7. [7] Ramalingam VV, Dandapath A, Raja MK. Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering & Technology. 2018 Mar
8. [8] Krishnaiah V, Narsimha G, Chandra NS. Heart disease prediction system using data mining techniques and intelligent fuzzy approach: a review. international journal of computer applications. 2016
9. [9] Patel J, TejalUpadhyay D, Patel S. Heart disease prediction using machine learning and data mining technique. Heart Disease. 2015
10.[10] Dangare CS, Apte SS. Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications. 2012 Jun;47