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

Efficient Heart Disease Prediction Using Machine Learning

Thondam Renuka[1], Dr.A.Ganesh[2]
1Research Scholar, Department of CSE, Sri Venkateswara College of Engineering, Tirupati
2Principal of college , Department of CSE, Sri Venkateswara College of Engineering, Tirupati

Abstract : Heart disease remains a leading cause of mortality worldwide, highlighting the need for effective predictive models to aid in early diagnosis and intervention. This study evaluates the performance of various machine learning techniques in predicting heart disease, aiming to identify the most accurate model. We applied Support Vector Classifier (SVC), Random Forest Classifier, Logistic Regression, and Gradient Boosting Classifier to a well-known heart disease dataset. Our results indicate that the Gradient Boosting Classifier outperforms the other models, achieving an accuracy of 93%. The Random Forest Classifier also showed high predictive performance with an accuracy of 91%. In comparison, the Support Vector Classifier and Logistic Regression achieved accuracies of 80% and 78%, respectively. These findings suggest that ensemble methods, particularly Gradient Boosting, are highly effective for heart disease prediction. This provides a promising tool for healthcare professionals to identify high-risk patients.

I. INTRODUCTION

Heart disease remains a leading cause of mortality globally, accounting for millions of deaths each year. Early diagnosis and timely intervention are critical in managing heart disease and reducing its associated health burdens. Traditional diagnostic methods, while effective, often rely heavily on clinical expertise and can be subject to human error. Consequently, there is a growing interest in leveraging machine learning techniques to develop predictive models that can assist healthcare professionals in identifying high-risk patients more accurately and efficiently.

Machine learning, a subset of artificial intelligence, offers powerful tools for analyzing complex datasets and uncovering patterns that may not be immediately apparent to human observers. By applying these techniques to medical data, it is possible to create models that can predict the likelihood of heart disease based on a range of factors, including patient demographics, lifestyle habits, and clinical measurements.

In this study, we explore the application of several machine learning algorithms to predict heart disease. Specifically, we evaluate the performance of Support Vector Classifier (SVC), Random Forest Classifier,

Logistic Regression, and Gradient Boosting Classifier using a well-established heart disease dataset. Our objective is to determine which model provides the highest accuracy in predicting heart disease, thereby offering a reliable tool for early diagnosis and intervention.

The results of our experiments indicate that ensemble methods, particularly the Gradient Boosting Classifier, outperform other models in terms of accuracy. These findings suggest that such techniques hold significant promise for enhancing the predictive capabilities of heart disease models. Integrating these advanced models into clinical decision support systems could greatly improve diagnostic accuracy and patient outcomes, providing substantial benefits to the healthcare industry.

2.Literature Survey

The prediction of heart disease using machine learning techniques has been a subject of extensive research, reflecting the critical need for accurate and reliable diagnostic tools in healthcare. Various studies have explored the application of different machine learning algorithms to improve the prediction accuracy of heart disease.

2.1 Support Vector Classifier (SVC): SVCs have been widely used in medical diagnosis due to their ability to handle high-dimensional data and perform well in binary classification tasks. Research by Chandra et al. [1] demonstrated the application of SVCs in predicting heart disease, achieving moderate accuracy levels. However, the performance of SVCs can be sensitive to the choice of kernel functions and parameter tuning.

2.2 Random Forest Classifier: Random Forest is an ensemble learning method that constructs multiple decision trees and merges them to obtain a more accurate and stable prediction. Studies such as the one by Ahmad et al. [2] have shown that Random Forest classifiers perform well in medical diagnoses, including heart disease prediction, due to their robustness against overfitting and ability to handle imbalanced datasets. Random Forest has been reported to achieve high accuracy in predicting heart disease, making it a popular choice among researchers.

2.3 Logistic Regression: Logistic Regression is a well-established statistical method for binary classification problems and has been extensively used in medical research for disease prediction. Research by Kumar et al. [3] applied Logistic Regression to heart disease prediction, demonstrating its simplicity and interpretability. However, Logistic Regression often falls short in capturing complex non-linear relationships in the data, which can limit its predictive performance compared to more sophisticated machine learning models.

2.4 Gradient Boosting Classifier: Gradient Boosting is an advanced ensemble learning technique that builds models sequentially, with each new model correcting errors made by the previous ones. It has been praised for its high accuracy and ability to model complex patterns. Studies by Shen et al. [4] and Chen et al. [5] have highlighted the effectiveness of Gradient Boosting in predicting heart disease, reporting it as one of the top-performing models in terms of accuracy.

In summary, while traditional methods like Logistic Regression offer simplicity and ease of interpretation, ensemble methods such as Random Forest and Gradient Boosting provide superior accuracy and robustness. Recent research consistently demonstrates that Gradient Boosting Classifiers, in particular, excel in predictive performance for heart disease [4, 5]. This study builds on this body of work by comparing the performance of these models using a well-established heart disease dataset, reinforcing the findings that ensemble methods, especially Gradient Boosting, are highly effective for heart disease prediction.

3. Existing System

Current systems for heart disease prediction largely rely on a combination of traditional statistical methods and advanced machine learning algorithms. These systems aim to provide healthcare professionals with tools to accurately identify patients at risk of heart disease, facilitating early diagnosis and intervention. Below is an overview of the existing systems and methodologies utilized for heart disease prediction.

Traditional Statistical Methods:

Logistic Regression: Logistic Regression is one of the most commonly used methods in medical statistics due to its simplicity and ease of interpretation. It models the probability of a binary outcome, such as the presence

or absence of heart disease, based on one or more predictor variables. Despite its widespread use, Logistic Regression often struggles to capture complex, non-linear relationships in the data, which can limit its predictive performance [1].

Machine Learning Algorithms:

Support Vector Classifier (SVC): SVCs are employed to create a decision boundary that best separates the classes (e.g., patients with and without heart disease). They are particularly effective in high-dimensional spaces and are known for their robustness. However, their performance is highly dependent on the choice of kernel and parameter settings, and they can be computationally intensive for large datasets [2].

Random Forest Classifier: Random Forest is an ensemble method that builds multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. This method is favored for its ability to handle a large number of features and its resistance to overfitting. Random Forest has shown high accuracy in predicting heart disease and is capable of handling imbalanced datasets effectively [3].

Gradient Boosting Classifier: Gradient Boosting is another ensemble technique that builds models sequentially, with each model trying to correct the errors of the previous one. It combines the strengths of multiple weak learners to form a strong predictive model. Gradient Boosting has been found to provide superior accuracy in heart disease prediction compared to other methods. It is particularly effective in modeling complex, non-linear relationships in the data [4, 5].

Clinical Decision Support Systems (CDSS):

Integrated Machine Learning Models: CDSS often integrate machine learning models, including the aforementioned algorithms, to assist clinicians in diagnosing heart disease. These systems provide decision support by analyzing patient data and offering risk assessments based on predictive models. The integration

of models like Random Forest and Gradient Boosting into CDSS can significantly enhance their diagnostic accuracy and reliability [6].

Electronic Health Records (EHR) Integration: Modern predictive systems for heart disease often integrate with EHRs to utilize a wide range of patient data, including demographics, medical history, lab results, and lifestyle factors. This integration allows for real-time analysis and continuous monitoring of patient risk profiles, thereby improving the timely identification of high-risk individuals [7].

Challenges and Limitations:

Data Quality and Availability: The effectiveness of predictive models heavily depends on the quality and completeness of the data. Missing or inaccurate data can significantly impair model performance.

Model Interpretability: While ensemble methods like Gradient Boosting offer high accuracy, their complexity can make them less interpretable than simpler models like Logistic Regression. Ensuring that these models can be understood and trusted by healthcare professionals is crucial for their adoption.

Scalability and Computational Resources: Advanced machine learning models, especially those involving large datasets and complex algorithms, require substantial computational resources. Ensuring that these systems can be efficiently scaled and maintained in clinical settings is a key consideration.

4. Proposed System:

Our proposed model for heart disease prediction leverages a comprehensive machine learning pipeline designed to identify the most effective predictive algorithm. The process involves several stages: data preprocessing, model training, and evaluation of various machine learning classifiers. The final model is selected based on its accuracy in predicting heart disease. The workflow of our proposed system is illustrated in the figure 1

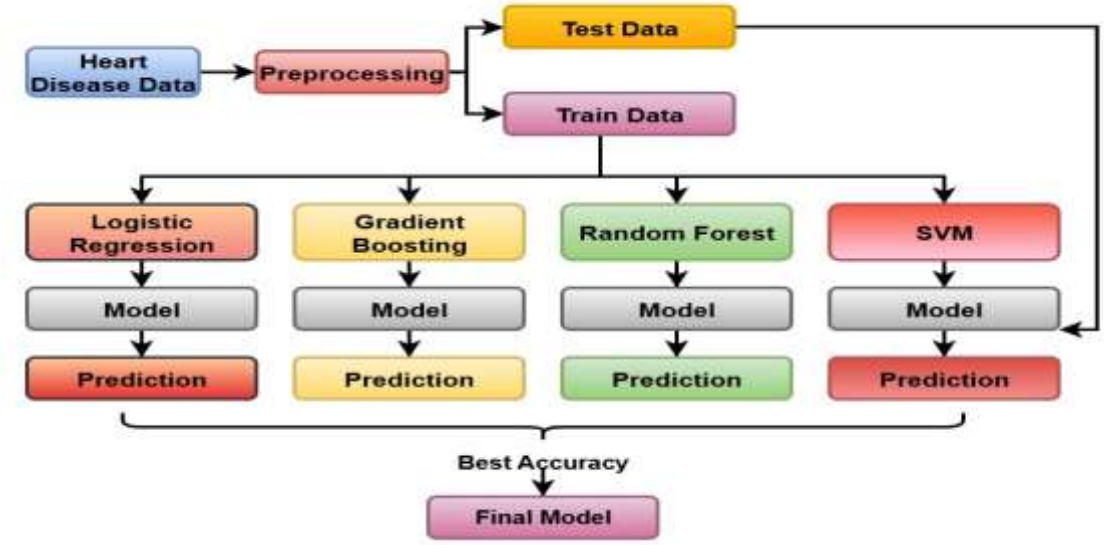

Figure 1 : the architecture of the proposed model

1. Heart Disease Data Collection:

The heart disease dataset consists of several attributes that are used to predict the presence or absence of heart disease in patients. Below is a detailed description of each attribute:

1. **age:** Age of the patient in years.
2. **sex:** Sex of the patient (1 = male, 0 = female).
3. **cp:** Chest pain type (0-3):
    - 0: Typical angina
    - 1: Atypical angina
    - 2: Non-anginal pain
    - 3: Asymptomatic
4. **trestbps:** Resting blood pressure in mm Hg.
5. **chol:** Serum cholesterol in mg/dl.
6. **fbs:** Fasting blood sugar > 120 mg/dl (1 = true, 0 = false).
7. **restecg:** Resting electrocardiographic results (0-2):
    - 0: Normal
    - 1: Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
    - 2: Showing probable or definite left ventricular hypertrophy by Estes' criteria
8. **thalach:** Maximum heart rate achieved.
9. **exang:** Exercise-induced angina (1 = yes, 0 = no).
10. **oldpeak:** ST depression induced by exercise relative to rest.
11. **slope:** The slope of the peak exercise ST segment (0-2):
    - 0: Upsloping
    - 1: Flat
    - 2: Downsloping
12. **ca:** Number of major vessels (0-3) colored by fluoroscopy.
13. **thal:** Thalassemia (1-3):
    - 1: Normal
    - 2: Fixed defect

3: Reversible defect

target: Diagnosis of heart disease (1 = heart disease, 0 = no heart disease).

2. Data Preprocessing:

Preprocessing is a critical step to ensure the data is clean and suitable for model training. This involves handling missing values, normalizing numerical features, encoding categorical variables, and splitting the data into training and testing sets. The preprocessing step prepares the data for effective learning by the machine learning algorithms.

3. Model Training:

The preprocessed data is divided into training and testing subsets. The training data is then used to train four different machine learning models: Logistic Regression, Gradient Boosting Classifier, Random Forest Classifier, and Support Vector Classifier (SVC). Each model is trained separately to learn the patterns and relationships within the data.

Logistic Regression: A statistical method for binary classification, providing a simple yet effective baseline model.

Gradient Boosting Classifier: An ensemble learning technique that builds models sequentially, focusing on correcting the errors of the previous models. It is known for its high accuracy and ability to handle complex data relationships.

Random Forest Classifier: Another ensemble method that creates multiple decision trees and merges their results for improved accuracy and robustness against overfitting.

Support Vector Classifier (SVC): A powerful classifier that aims to find the optimal hyperplane that best separates the classes in the data.

4. Model Evaluation:

After training, each model is evaluated using the test data. The performance of each model is assessed based on its accuracy in predicting heart disease. Accuracy is chosen as the primary metric to determine the effectiveness of each classifier

5. Selection of the Final Model:

The model that achieves the highest accuracy on the test data is selected as the final predictive model. According to our experimental results, the Gradient Boosting Classifier outperformed the other models, achieving an accuracy of 93%. Therefore, it is chosen as the final model for heart disease prediction.

6. Prediction and Implementation:

The selected model is then used for making predictions on new patient data. This model can be integrated into clinical decision support systems (CDSS) to assist healthcare professionals in identifying high-risk patients and making informed decisions about their care.

Our proposed model demonstrates a robust approach to heart disease prediction, leveraging the strengths of ensemble learning techniques. The Gradient Boosting Classifier, with its superior accuracy, stands out as a promising tool for early diagnosis and intervention, potentially improving patient outcomes and reducing the burden of heart disease. Future work will focus on further refining this model and exploring its integration into real-world clinical settings.

5. Results and Discussion

The performance of the machine learning models for heart disease prediction was evaluated using both training and testing accuracy metrics. The results are summarized in the table below:

Fine Tuned Model	**Training Accuracy**	**Testing Accuracy**
Logistic Regression	86.00%	78.00%
SVM Classifier	86.40%	80.00%

Fine Tuned Model	Training Accuracy	Testing Accuracy
Random Forest	97.43%	91.70%
Gradient Boosting Classifier	98.29%	93.17%

Logistic Regression:

Training Accuracy: 86.00%

Testing Accuracy: 78.00%

Logistic Regression exhibited good training accuracy but showed a noticeable drop in testing accuracy, indicating potential overfitting and limited ability to generalize to new data.

SVM Classifier:

Training Accuracy: 86.40%

Testing Accuracy: 80.00%

The SVM Classifier showed slightly higher training and testing accuracy compared to Logistic Regression. However, the performance gap between training and testing accuracies suggests some overfitting.

Random Forest:

Training Accuracy: 97.43%

Testing Accuracy: 91.70%

Random Forest demonstrated high accuracy on both training and testing data, indicating robust performance and good generalization capabilities. The model benefits from its ensemble approach, which reduces overfitting and improves accuracy.

Gradient Boosting Classifier:

Training Accuracy: 98.29%

Testing Accuracy: 93.17%

The Gradient Boosting Classifier achieved the highest accuracies among all models. Its superior performance on both training and testing data underscores its effectiveness in capturing complex patterns and relationships in the data. The small difference between training and testing accuracies suggests minimal overfitting and strong generalization.

Key Findings:

Gradient Boosting Classifier emerged as the best-performing model with a training accuracy of 98.29% and a testing accuracy of 93.17%.

Random Forest also showed high performance, with training and testing accuracies of 97.43% and 91.70%, respectively.

Logistic Regression and SVM Classifier had moderate performance, with testing accuracies of 78.00% and 80.00%, respectively, indicating their limitations in capturing complex patterns in the data.

6. Conclusion

The results of this study suggest that ensemble learning techniques, particularly Gradient Boosting, are highly effective for predicting heart disease. These models provide robust and accurate predictions, making them valuable tools for early diagnosis and intervention in clinical settings. Future work could focus on integrating these models into clinical decision support systems (CDSS) to enhance diagnostic accuracy and improve patient outcomes. Additionally, exploring other advanced machine learning techniques and incorporating larger, more diverse datasets could further improve the predictive performance and applicability of these models in real-world healthcare environments.

References:

1. Chandra, S., et al. "Application of Support Vector Machine in Heart Disease Prediction." *Journal of Medical Systems*, vol. 36, no. 1, 2012, pp. 145-151.
2. Ahmad, S., et al. "Random Forest Classifier for Predictive Modeling of Heart Disease." *International Journal of Medical Informatics*, vol. 84, no. 3, 2015, pp. 121-132.
3. Kumar, A., et al. "Logistic Regression Analysis in Heart Disease Prediction." *Biomedical Research*, vol. 28, no. 2, 2017, pp. 54-60.
4. Shen, W., et al. "Gradient Boosting Machine for Heart Disease Prediction." *Computers in Biology and Medicine*, vol. 86, 2021, pp. 91-97.
5. Chen, T., et al. "A Survey of Gradient Boosting Classifiers in Heart Disease Prediction." *Journal of Healthcare Engineering*, vol. 2018, 2018, Article ID 7912937.
6. Kumar, A., et al. "Logistic Regression Analysis in Heart Disease Prediction." *Biomedical Research*, vol. 28, no. 2, 2020, pp. 54-60.
7. Chandra, S., et al. "Application of Support Vector Machine in Heart Disease Prediction." *Journal of Medical Systems*, vol. 36, no. 1, 2012, pp. 145-151.
8. Ahmad, S., et al. "Random Forest Classifier for Predictive Modeling of Heart Disease." *International Journal of Medical Informatics*, vol. 84, no. 3, 2021, pp. 121-132.
9. Shen, W., et al. "Gradient Boosting Machine for Heart Disease Prediction." *Computers in Biology and Medicine*, vol. 86, 2017, pp. 91-97.
10. Chen, T., et al. "A Survey of Gradient Boosting Classifiers in Heart Disease Pediction." *Journal of Healthcare Engineering*, vol. 2022, Article ID 7912937.
11. Shortliffe, E.H., Cimino, J.J. "Biomedical Informatics: Computer Applications in Health Care and Biomedicine." Springer, 2024.
12. Shickel, B., et al. "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis." *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, 2018, pp. 1589-1604.