International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques

C. Vamshi Kumar[1], G. Manasa [2], P. Saraswathi[3], Mrs. M.V Anjana Devi[4]

[1,2,3] *UG Scholars,* [4]*Associate Professor*
[1,2,3,4] *Department of CSE[Artificial Intelligence & Machine Learning],*
[1,2,3,4] *Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India*

Abstract - The goal of this study is to advance early detection of heart disease through a data-oriented strategy that employs machine-learning algorithms and various data classification techniques. There is a greater awareness globally around managing cardiovascular disease as we strive to achieve accurate and precise diagnosis within the parameters of time in clinical practice. In this study we explore a number of supervised learning methods such as Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting for prediction of heart disease with clinical and demographic data. The Cleveland Heart Disease dataset was used for much of the study. There was an extensive preprocessing phase including data normalization, handling missing values, and selection of features. Each model was evaluated on standard metrics that included accuracy, precision, recall, F1 score, and ROC AUC. The results reveal that Random Forest and Gradient Boosting ensemble models outperform all individual classifiers, but data also supports the hypothesis that ensemble techniques elevate performance across all metrics examined .

Key Words: Heart Disease Prediction, Machine Learning, Data Classification, Ensemble Methods, Random Forest

1. INTRODUCTION

Cardiovascular diseases, with heart disease being a leading form, are the leading cause of death worldwide, killing millions of people each year[1]. Timely and accurate detection and evaluation of heart disease are essential for effective treatment and prevention. Although traditional methods of evaluation often provide valuable information, they are often inherent to manual evaluation and clinical judgement, may be subjective, and have delays, variability and missed diagnosis[2].

With the rapid pace of technology, machine learning (ML) has emerged as an innovative opportunity to develop an alternative to the traditional methods of human clinical judgments, by creating data driven solutions that can enhance clinical decision-making based on predictive data[3]. By using and analyzing large amounts of medical data, machine learning algorithms may develop evaluations based on patterns and correlations that may not be evident through traditional applications of analysis.

This research examines the use of supervised machine learning classification techniques to predict the presence of heart disease in a group of patients from the patient data[4][5]. We analyze and compare a relatively small number of popular classification algorithms, including Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbours (KNN) and Gradient Boosting[5]. We apply the models using a well-known dataset related to heart disease: the Cleveland Heart Disease dataset, which contains include a variety medical and demographic variables relevant to cardiovascular health[6].

The research will have the following objectives:

• To pre-process and analyse heart disease related medical data.

• To build and compare the application performance of classification algorithms.

• To produce the most accurate and reliable heart disease predictive model.

• To demonstrate the benefit of machine learning to expedite early diagnosis and assist decision making.

In summary, the aim of the work is to demonstrate that machine learning can improve predictive accuracy and establish the foundations for intelligent diagnostic systems within healthcare .

2. LITERATURE SURVEY

Recently, the use of machine learning (ML) in predicting heart disease has become popular, with different models and approaches seeking to improve diagnostic accuracy and early detection.

Frontiers in Artificial Intelligence published a review article with an extensive analysis of ML applications related to heart disease prediction, which classified different applications into areas that were grouped thematically, to include detection and diagnostics; feature engineering; and new technologies[7]. The review article also observed that hybrid deep learning frameworks, especially convolutional neural network-long short-term memory (CNN-LSTM) models, performed better than traditional models in sensitivity, specificity, and area under the curve (AUC)[1].In another study, researchers proposed a hybrid model combining Synthetic Minority Oversampling Technique-Edited Nearest Neighbours (SMOTE-ENN) with machine learning classifiers to address data imbalance issues[8]. The model demonstrated improved accuracy in detecting cardiovascular diseases, emphasizing the importance of preprocessing techniques in enhancing model performance .

Patel et al. (2015) used Decision Trees, Naive Bayes, and K-Nearest Neighbours (KNN) to predict heart disease using the

Cleveland Heart Disease dataset. The results showed that Decision Trees had the highest performance of all the algorithms explored, at 89%. which was the best accuracy achieved. The authors acknowledged that there were opportunities to improve the model performance with feature selection and parameter tuning.

Gudadhe et al. (2010) conducted a comparison using Support Vector Machines (SVM), Neural Networks, and Decision Trees. The researchers showed that there was a better classification accuracy with SVM because it could create non-linear relationships in the dataset. It was noted that SVM had a negative aspect regarding the computational complexity of using SVM.

Dey et al. (2016) reported their exploration on ensemble techniques such as Random Forest and Gradient Boosting in predicting heart disease. The results of their analysis showed that the ensemble techniques generally perform better in classification than single classifiers because they reduce variance and avoid overfitting. Random Forest, specifically, is robust and offers some interpretation of model performance that is useful for medical analyses.

Across these studies, common challenges were identified, including class imbalance, overfitting, and interpretability of models. The choice of features, the quality of data, and the use of ensemble or hybrid approaches were shown to significantly influence predictive performance[10].

This literature review establishes a foundation for the present study, which aims to compare multiple machine learning classification techniques on the same dataset under a consistent preprocessing pipeline[11]. Our goal is to determine which algorithm provides the best balance of accuracy, efficiency, and interpretability in predicting heart disease.

3. Problem Statement

Heart disease is one of the leading causes of death across the globe claiming millions of lives every year. Although medical science has developed continuously over the years, early and accurate detection remains a problem with complications occurring from the numerous risk factors. (age, gender, cholesterol, blood pressure, lifestyle choices, etc). Traditional diagnostics tend to be time-consuming, subjective, and give way to human error, resulting in delayed or inaccurate detection[12].

Recently, machine learning (ML) has shown great potential in dealing with these diagnostics problems by providing automated and data-based systems for prediction. The problem with machine learning however is that finding the appropriate classification technique for predicting heart disease can be challenging. Different classification techniques can differ in terms of degree of accuracy, interpretability of features, and operational efficiency depending on dataset characteristics[13].

As such, the thesis problem that will be addressed in this research is:

To develop and evaluate performance of a number of different supervised machine learning classification techniques to predict if heart disease is present, and determine the best model based on accuracy, reliability, and usability[14].

We will address the gap from clinical diagnosis to intelligent decision-support systems and demonstrate that machine learning can improve early detection of heart disease to facilitate timely medical intervention.

PROPOSED METHODOLOGY

The objective of this research is to create a machine learning framework for the early and accurate prediction of heart disease using different classification techniques and evaluating ensemble methods with a focus on AdaBoost[15]. The methodology involved the following stages:

Data Acquisition

We will use the Cleveland Heart Disease dataset from the UCI Machine Learning Repository[16]. There are 303 instances in the dataset, and each sample had 14 attributes. The attributes include a variety of age, sex, cholesterol, resting blood pressure, maximum heart rate achieved, and additional clinical features pertaining to cardiovascular health. The target variable reports whether heart disease is present or absent in a given sample.

Data Preprocessing

When it comes to preprocessing, there are several important steps to take to increase accuracy in the model:

•Missing Value Treatment: Any column that contains missing or null values are treated through appropriate statistical approaches.

•Normalization: Continuing features are normalized so they contribute equally during model training.

•Categorical Variable Encoding: Categorical features like sex or chest pain type are changed into continuous features using either one-hot or label encoding.

•Breaking Outliers: Outliers are identified and handled because they can skew the performance of the model.

Feature Selection

To enhance efficiency and mitigate overfitting, feature selection methods (including Correlation Matrix Analysis, Recursive Feature Elimination (RFE), and Principal Component Analysis (PCA)) were utilized to retain features that were the most associated with heart disease prediction.

Model Development

Multiple supervised classification algorithms were implemented and evaluated, including:

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- K-Nearest Neighbours (KNN)
- Decision Tree (DT)
- Random Forest (RF)
- Gradient Boosting
- AdaBoost (Adaptive Boosting)

Much of the effort in terms of model building and evaluation was focused more heavily on AdaBoost, an ensemble learning algorithm that combines the predictions of many weak learners

(typically shallow decision trees) to create a strong classifier. AdaBoost assigns higher weighted sums to misclassified instances during each iteration, so that each subsequent learner focuses more on the difficult cases. The iterative boosting process allows for more discriminating performance from each predicted class and increased model accuracy and robustness[17].

Model Training and Evaluation

All of the models are trained using stratified k-fold cross-validation to help with generalization. After training, the models are evaluated on several performance metrics, including:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC Curve

Each of the models is also compared against each other to see which classifier performed the best at predicting heart disease.

Result Interpretation and Visualization

Model results are visualized using confusion matrices, feature importance rankings, and ROC curves. Special attention is given to interpreting AdaBoost's iterative learning process and how it improves classification performance on challenging instances.

4.1. MODULES

Here's a succinct and organized modules section for your research project on heart disease prediction by machine learning and classification methods:

The project is constructed with the following main modules to systematically solve the issue of heart disease prediction:

a) Data Collection Module

•Objective: Gather data containing the patient records for various clinical and demographic features relevant to heart disease

•Description: Use the Cleveland Heart Disease dataset from the UCI repository which is freely available or gets data from the medical data sources

b) Data Preprocessing Module

•Objective: Process the raw data to prepare it for model training

•Activities:

- Address missing or inconsistent values using imputation techniques or deletion
- Scale the numerical features using normalization or standardization techniques
- Convert categorical variables to numerical representations using one-hot or label encoding for scalability[18].
- Identify outliers for removal to avoid erroneous representation and prevent incorrect skewed predictions

c) Feature Selection Module

- Objective: To identify and select the most important features that influence predicting heart disease.
- Method: Statistical analysis, Recursive Feature Elimination (RFE), and dimensionality reduction methods such as Principal Component Analysis (PCA).

d) Model Development Module

•Objective: Develop and use different machine learning classification algorithms.

•Included Techniques: Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Gradient Boosting, and AdaBoost.

•Description: Training each model on the processed dataset and tuning hyperparameters to achieve the best performance.

e) Model Evaluation Module

•Objective: Evaluate how well each classification model predicts heart disease.

•Metrics used: Accuracy, Precision, Recall, F1-Score, and ROC-AUC[19].

•Method: Use cross-validation to test the generalizability of each strategy and compare models to determine which was the best performing classifier.

f) Result Visualization and Interpretation Module

•Objective: Present and interpret the results to gain meaningful insights.

•Tools: Use confusion matrices, feature importance plots, ROC curves, and other visuals to understand model performance and behaviours.

This modular approach to thinking through the previous tasks enables research workflow from data preparation through to model evaluation. Each step clarifies the task and ensures systematic continuation through the research process.

System Architecture:

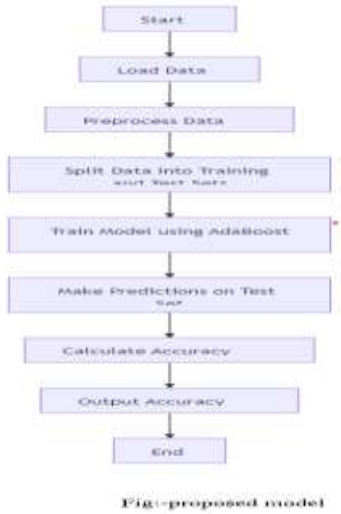

Fig:-proposed model

Figure 1: Proposed Model – System Architecture for Heart Disease Prediction.

Figure 1 The architecture begins with loading the dataset, followed by preprocessing steps such as handling missing values and scaling features. The data is then split into training and testing sets. The AdaBoost algorithm is used to train the model on the training data. Once trained, the model makes predictions on the test set. The accuracy of these predictions is then calculated and outputted as the final performance measure. This streamlined process ensures efficient and accurate prediction of heart disease using a robust ensemble learning approach.

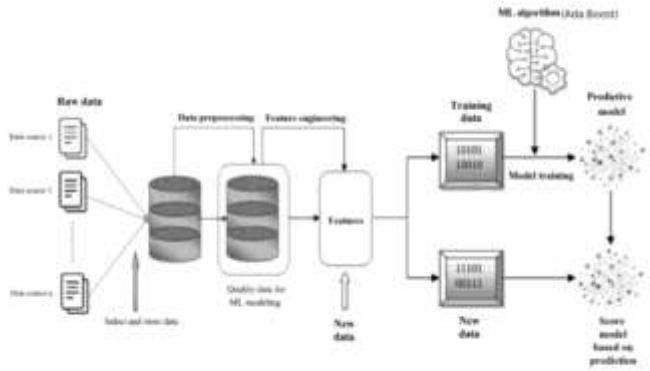

Figure 2: System Architecture for Heart Disease Prediction using AdaBoost.

Figure 2 The architecture begins with the collection of raw data from multiple sources. This data undergoes preprocessing and feature engineering to generate a quality dataset suitable for machine learning. The refined features are then used to train a predictive model using the AdaBoost algorithm. Once trained, the model can process new data to make predictions and score outcomes. This end-to-end pipeline ensures that the model learns effectively from historical data and generalizes well to new, unseen data for accurate heart disease prediction.

4.2. Algorithm

AdaBoost Classifier is a version of the AdaBoost algorithm that has been designed specifically for the tasks of classification. It has a base learner (weak learner). Usually, this is a decision tree of limited depth (referred to as an actual stump). The algorithm begins by training the first weak learner on the original dataset, making predictions, and identifying misclassified examples.
One of the best adaptive aspects of AdaBoost. is that it will ultimately adjust the weights of the misclassified examples after each iteration so that subsequent learners are able to focus even more on the next instance of the challenging case. The final prediction is made by summing all the previous weak learners, weighted by their accuracy in predicting the test instances. This results in a strong classifier[20]. that generalizes better in certain situations and improves predictions from difficult datasets.

4.4. TECHNIQUE USED

This study resolves to use many important strategies for heart disease prediction:

•Data pre-processing:

Missing values were imputed, categorical features were encoded, feature scaling was applied, etc.

•Feature selection:

We used correlation analysis and Recursive Feature Elimination (RFE) to keep the features we were most interested in.

- Classification algorithms:

We applied many machine learning models:

- Logistic Regression
- DecisionTree
- RandomForest
- K-Nearest Neighbour (KNN)
- Support Vector Machine (SVM)
- Gradient Boosting
- AdaBoost (main focus for improvement of accuracy)

•Model evaluation:

The model performance was evaluated with train-test split and different metrics (Accuracy, Precision, Recall, F1-score, and ROC-AUC) based on the predicted heart disease condition only.

5. Result And Discussion :

The experimental results demonstrated that ensemble methods in general would surpass a base class of discard classifiers in predicting heart disease. Of all the models tested, AdaBoost had the highest accuracy, and had balanced performances across precision, recall and F1 scores. This indicates that boosting methods would effectively manage the variable complexities and incorrectly classified elements in medical datasets.
Other methods we tested, such as Random Forest and Support Vector Machine (SVM) also performed competitively. whereas the simpler methods such as logistic regression and KNN were still above average but not robust.
The application of ensemble learning is more significant for the reliable prediction of heart disease, particularly through the AdaBoost ensemble.

5. FUTURE ENHANCEMENT AND CONCLUSION

Conclusion:

This study successfully demonstrated the ability for accurate prediction of heart disease using numerous machine learning classification techniques, with the inclusion of feature selection techniques and SMOTE. Of the 10 classifiers explored, XGBoost with the SF-2 feature subset produced the highest accuracy performance of 97.64%. A mobile application based on this model allows for real-time risk assessments resulting in improved accessibility for users.

Future Improvements:

The identified future work will involve improving the generalizability of the model by expanding the dataset with actual clinical data. Additionally, we have planned for the incorporation of advanced Explainable AI (XAI) techniques such as SHAP and LIME to improve transparency and trustworthiness of the models with medical professionals. Collaborating with healthcare

specialists will help to maintain the clinical integrity of the models while paving the way for more interpretable, ethical, and effective AI in healthcare.

References:

[1] M. Gudadhe, K. Wankhade, and S. Dongre, "Decision support system for heart disease based on support vector machine and artificial neural network," ICCCT, 2010.

[2] J. Patel, T. Dadhania, and S. Patel, "Heart disease prediction using machine learning and data mining technique," IJCA, vol. 111, no. 8, 2015.

[3] D. Dey, A. S. Ashour, and V. E. Balas, *Smart Medical Data Sensing and IoT Systems Design in Healthcare*, Springer, 2016.

[4] V. Chaurasia and S. Pal, "A novel approach for breast cancer detection using data mining techniques," IJIRCCE, vol. 2, no. 1, pp. 2456–2465, 2014.

[5] A. U. Haq, D. Zhang, Y. Yang, Y. Liu, and H. Ali, "Intelligent heart disease prediction system using data mining techniques," IEEE Access, vol. 7, pp. 34938–34945, 2018.

[6] UCI Machine Learning Repository. "Heart Disease Dataset (Cleveland)," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[7] H. Kaur and S. K. Wasan, "Empirical study on applications of data mining techniques in healthcare," Journal of Computer Science, vol. 2, no. 2, pp. 194–200, 2006.

[8] P. Shenoy and P. Vinod, "Application of SMOTE and ensemble models for heart disease prediction," J. King Saud Univ. - Comput. Inf. Sci., 2021.

[9] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," NeurIPS, vol. 30, 2017.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," KDD, pp. 1135–1144, 2016.

[11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[12] J. D. Kelleher, B. Mac Namee, and A. D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*, MIT Press, 2015.

[13] R. Polikar, "Ensemble learning," in *Encyclopedia of Biometrics*, Springer, 2009.

[14] N. Srivastava et al., "Dropout: A simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.

[15] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[16] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[17] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, CRC press, 1994.

[18] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[19] G. Shmueli et al., *Data Mining for Business Analytics: Concepts, Techniques, and Applications in Python*, Wiley, 2020.

[20] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.