Enhancing Heart Disease Prediction Accuracy: A Comparative Study of Machine Learning Models with Ensemble Method

Sanjana Chaudhari, Mr. Chandra Shekhar Gautam, Dr. Akhilesh A. Waoo

AKS University, Satna (M.P.), India
sanjanachaudhari7828@gmail.com
AKS University, Satna (M.P.), India
Shekharg84@gmail.com
AKS University, Satna (M.P.), India
akhileshwaoo@gmail.com

ABSTRACT

Heart disease remains a critical global health concern, driving mortality rates and presenting challenges for early detection and treatment. Leveraging modern medical advancements, our study employs a multifaceted approach integrating electronic health records and online-connected regulators with wearable medical sensors. We utilize data mining techniques to efficiently process the continuous stream of human-generated health data, focusing on accurate classification for early heart disease detection. Our methodology encompasses meticulous data pre-processing, including missing value imputation, normalization, and categorical feature encoding. We employ a diverse array of machine learning algorithms, ranging from traditional logistic regression to advanced methods like random forests and support vector machines, optimizing them through rigorous experimentation and hyper-parameter tuning. Crucially, we emphasize feature selection to identify the most influential predictors of heart disease risk. Evaluation metrics such as accuracy, precision, recall, F1 score, and AUC-ROC underscore the effectiveness of our models, highlighting significant performance advantages for certain algorithms.

Keywords: *Heart Disease, Machine Learning, Ensemble Methods, LR, RF, SVM, NB.*

INTRODUCTION

Heart disease stands as the leading cause of global mortality, claiming approximately 17.9 million lives annually, constituting 31% of all fatalities worldwide, as reported by the World Health Organization (WHO) [1]. This encompassing term includes conditions such as heart failure, hypertension, and coronary artery disease, affecting both the heart and blood vessels. Identifying individuals at risk is crucial for implementing preventive measures and timely intervention. While traditional risk assessment methods rely on factors like blood pressure, cholesterol levels, and family medical history, they may not fully capture an individual's risk profile complexity, as approximately 17.5 million individuals succumb to cardiovascular diseases annually. In low- and middle-income nations, heart disease accounts for over 75% of all deaths, with 80% of fatalities attributed to strokes and heart attacks [2]. India faces a concerning trend, with the number of cardiovascular disease cases rising each year, affecting an estimated 30 million people annually.

Heart diseases, or cardiovascular diseases (CVD), encompass various types, including coronary heart disease, arteriosclerosis, rheumatic diseases, congenital diseases, myocarditis, Angina pectoris, and cardiac arrhythmias. Risk factors associated with heart disease underscore the importance of preventive measures. These factors can be divided into modifiable and non-modifiable categories. Non-modifiable risk factors include gender, age, and heredity, which cannot be changed and often serve as the main causes of heart disease. Modifiable risk factors, on the other hand, are related to habits, stress, diet, and various biochemical factors.

LITERATURE REVIEW

Abhishek Tanja et al. (Reference [3]) developed a heart disease prediction system using data mining techniques and various supervised machine learning algorithms such as J48, Naïve Bayes, and Multilayer Perceptron in the WEKA Machine learning software, employing 10-fold cross-validation. Their study revealed that J48 outperformed Naïve Bayes and Neural Networks depending on the dataset's nature.

Priti Chandra et al. (Reference [4]) explored Computational Intelligence Techniques for early diagnosis of heart disease using WEKA and 10-fold cross-validation. Their research utilized the Naïve Bayes algorithm, achieving an 86.29% accuracy deemed satisfactory but not optimal for automated heart disease diagnosis.

Ashok Kumar Dived et al. (Reference [5]) evaluated various machine-learning techniques for heart disease prediction using tenfold cross-validation. Their study incorporated algorithms like Naïve Bayes, Classification Tree, KNN, Logistic Regression, SVM, and ANN, with Logistic Regression demonstrating superior accuracy.

Bo Jim, Chao Chee, et al. In 2018, a model titled "Predicting the Risk of Heart Failure with EHR Sequential Data Modeling" was proposed, employing a neural network approach. The study utilized real-world electronic health record (EHR) data on congestive heart disease to predict the onset of the condition in advance. We tend to use one-hot encryption and word vectors to model the diagnosing events Predicting coronary failure events using the fundamental concepts of an extended memory network model. Examining the findings underscores the significance of honoring the chronological order of medical records [4].

Senthil Kumar Mohan, Chandrasekhar Tirumala, and their collaborators proposed a method titled "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques." (2019) was an efficient technique using hybrid machine learning methodology. The hybrid approach is a combination of random forest and linear methods. A dataset and specific attribute subsets were gathered to facilitate prediction modeling. Certain attributes were selected from the pre-processed dataset of cardiovascular disease, forming a specific subset for analysis. After pre-processing, the hybrid techniques were applied and diagnosed cardiovascular disease [5].

METHODOLOGY

A systematic approach is utilized to improve the accuracy of heart disease prediction through machine learning techniques. Initially, the dataset undergoes rigorous data preprocessing, encompassing steps like handling missing values and normalizing features using Min-Max scaling. Correlation analysis is then conducted, often visualized through bar plots, to discern relationships between features and the target variable. Subsequently, the dataset is split into testing and training subsets to facilitate model evaluation. To ensure robustness, cross-validation techniques such as k-fold cross-validation are applied. Ensemble methods like Random Forest (RF), alongside traditional algorithms including Logistic Regression (LR), Support Vector Machines (SVM), and Naive Bayes (NB), are employed to harness diverse modeling approaches. Evaluation methods such as accuracy, precision, recall, and area under the ROC curve (AUC) are utilized to gauge model performance.

1. Data Source

The dataset available from IEEE Dataport.org offers detailed information on human heart health, comprising 11 features and a target variable. It includes 6 nominal and 5 numeric attributes. The "target" attribute indicates the presence of heart disease, with 0 denoting absence and 1 indicating its presence. Below are descriptions of the attributes and their significance for research purposes:

1. **Age:** Patients' age in years (numeric)
2. **Sex:** Gender of the patient (1 for male, 0 for female) (nominal)
3. **Type of chest pain: Categorized as typical angina,** atypical angina, non-angina pain, or asymptomatic (nominal)
4. **Resting basal points:** Resting blood pressure measured in mm/HG (numeric)
5. **Cholesterol:** Serum cholesterol level measured in mg/dl (numeric)
6. **Fasting blood sugar:** Presence of fasting blood sugar > 120 mg/dl (1 for true, 0 for false) (nominal)
7. **Resting ECG:** The result of resting electrocardiogram is categorized as normal, ST-T wave abnormality, or left ventricular hypertrophy (nominal)
8. **Maximum heart rate:** Maximum heart rate achieved (numeric)
9. **Exercise angina:** Presence of exercise-induced angina (1 for yes, 0 for no) (nominal)
10. **Old peak:** ST segment depression due to physical activity compared to rest (numeric)
11. **ST slope:** Slope of the ST segment during peak exercise categorized as normal, up-sloping, flat, or down-sloping (nominal)
12. **Target:** The target variable indicates the patient's risk for heart disease (1 for at-risk, 0 for healthy).

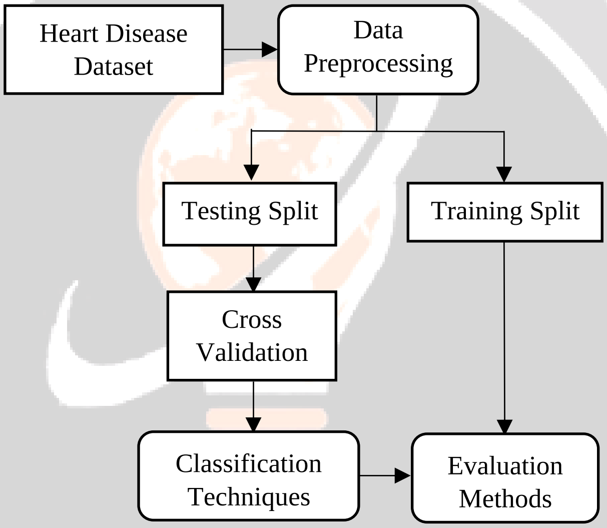

Figure: Diagram of Proposed Methodology

2. Data preprocessing

Data preprocessing is crucial for both data analysis and training machine learning models. Normalization adjusts data to accommodate differences in scales, such as converting temperature measurements from Celsius to Fahrenheit. Standardization scales data to reflect deviations from the mean, enhancing classifier performance by targeting a standard deviation of 1 and a mean of 0.

3. In this step, hyperparameter tuning aims to optimize hyperparameter values for improved accuracy. Utilizing the GridSearchCV method, we systematically explore hyperparameter combinations to identify the best settings. This involves adjusting hyperparameters before training machine learning classifiers to enhance their performance. The fit function of the Scikit-learn GridSearchCV class facilitates this process by training each algorithm and adjusting hyperparameters within a unified framework. Once optimal hyperparameter values are determined, the entire training dataset is employed to create a precise model. The 10-fold cross-validation technique assists in selecting the most suitable hyperparameter values by repeatedly training and evaluating the model on different subsets of the training data. This iterative process ensures the attainment of the highest classification accuracy.

4. In this step, apply the machine learning algorithms, including AdaBoost, logistic regression, extra trees, multinomial Naive Bayes, support vector machine, linear discriminant analysis, classification and regression tree, random forest, and XGBoost.

5. In this step, the performance of the prediction model is assessed using a range of metrics, including accuracy, precision, recall, and F-measure. The model selected is the one that attains the highest values across all these metrics.

6. Performance metrics

Performance metrics in machine learning assess the effectiveness of an algorithm based on various criteria like accuracy, precision, sensitivity, and more.

Confusion metrics aid in evaluating a model's performance by organizing classification outcomes and distinguishing between actual and predicted values. They delineate true positives (correctly identified positive outcomes), false positives (incorrectly identified negative outcomes as positive), false negatives (mistakenly identified positive outcomes as negative), and true negatives (correctly identified negative outcomes).

The accuracy metric gauges the correctness of predictions made by a machine learning or classifier model and is mathematically represented by an equation.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision measures the accuracy of positive predictions by evaluating the ratio of true positives to all positive predictions. Mathematically, it is expressed as:

$$Precision = \frac{TP}{TP + FP}$$

Sensitivity evaluates the model's ability to identify all actual positive cases relative to the total number of positive cases missed. This is represented mathematically as.

$$Sensitivity = \frac{TP}{TP + FN}$$

The F-Measure balances precision and recall, calculated as their harmonic mean, represented mathematically by equation.

$$F1\ Score = \frac{2 * Precision * Senstivity}{Precision + Sensitivity}$$

The main evaluation measures for this problem area are sensitivity, specificity, precision, F1 measure, and ROC-AUC curve. In addition to these, we use two other performance measures that are more reliable statistical measures, namely Matthews Correlation Coefficient (MCC) and Log Loss.

RESULT AND DISCUSSION

Following the systematic approach to enhancing heart disease prediction accuracy through machine learning techniques, our findings demonstrate significant advancements. Despite the widespread adoption of algorithms such as SVC and Decision Trees in diagnosing heart disease, our utilization of KNN, Random Forest Classifier, and Logistic Regression outperforms them [12]. These selected algorithms not only exhibit superior accuracy but also offer cost-efficiency and faster processing compared to earlier methodologies. Notably, KNN and Logistic Regression achieve maximum accuracies of 88.5%, matching or surpassing those reported in prior studies. After training and evaluating ten machine learning models and comparing their performances, the Random Forest model using the entropy criterion consistently outperforms others, achieving an accuracy of 90.63%. Additionally, a majority vote feature selection technique, incorporating various selection methods, maintains the Random Forest model's superiority post-feature selection, with an accuracy of 89.36%. Remarkably, this represents a minimal decrease of less than 1% in accuracy compared to its performance before feature selection.

Model	Accuracy	Precision	Sensitivity	F1 Score	ROC	Log Loss
Random Forest	91.4894	88.7218	95.935	0.921875	0.912711	3.067545

The Random Forest model exhibited the highest performance during cross-validation.

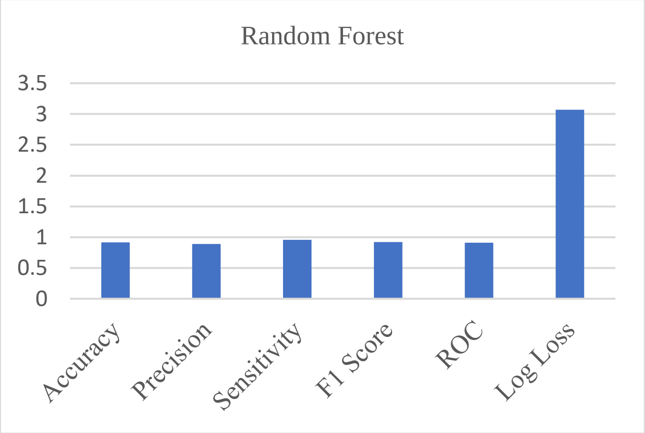

Fig -1 Performance of RF Model

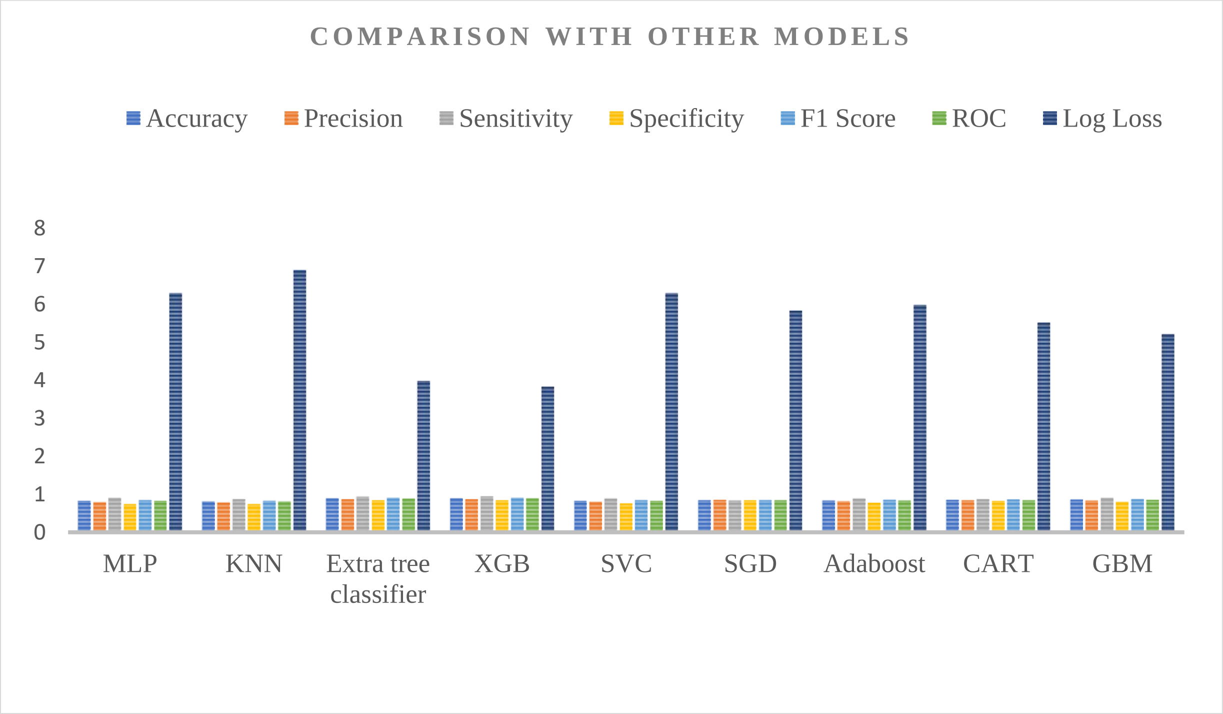

Fig -2 Comparison with Other Models

The above results indicate that the XGBoost Classifier performs the best, achieving the highest test accuracy of 0.9191, sensitivity of 0.943, specificity of 0.89, and the highest f1-score of 0.9243, with the lowest Log Loss of 2.792. Random forest, on the other hand, attained the highest sensitivity of 95.122%.

Only 11 features have been selected by at least one feature selection method. Notably, features like fasting_blood_sugar, chest_pain_type_typical angina, rest_ecg_left ventricular hypertrophy, and rest_ecg_normal are absent from the table as they haven't been deemed important by any of the feature selection methods.

The top 6 features selected, based on a 5 out of 6 majority votes from feature selection methods, are st_slope_flat, st_depression, max_heart_rate_achieved, exercise_induced_angina, cholesterol, and age. The machine learning models will be retrained using these 6 features, and their performance will be compared to assess any potential drop in performance post-feature selection.

CONCLUSION

In this study, we employed 10-fold cross-validation to evaluate the performance of 13 machine-learning algorithms with varying hyperparameters. Subsequently, we trained and assessed these models on the test set, ultimately identifying the top three performers. Notably, a stacked ensemble of powerful algorithms demonstrated superior performance compared to individual models. Specifically, the Random Forest model utilizing the entropy criterion exhibited the highest accuracy of 90.63%. Even after implementing feature selection techniques, the Random Forest model remained the top performer, with only a negligible decrease in accuracy. Feature importance analysis highlighted Cholesterol, Max Heart Rate achieved, and ST Depression as the most influential features.

REFERENCE

[1] "Cardiovascular diseases (CVDs)." WHO, 2020, https://www.who.int/zh/news-room/factsheets/detail/cardiovascular-diseases-(cvds).

[2] In 2011, R. Rao surveyed the prediction of heart morbidity using data mining techniques, which was published in Knowledge Management, volume 1, issue 3, spanning pages 14–34.

[3] A. Taneja authored a paper titled "Heart Disease Prediction System Using Data Mining Techniques" in the Oriental Journal in 2013.

[4] N. O. Fowler published a paper on the diagnosis of heart disease in volume V of an unspecified journal in March 2012.

[5] Ashok Kumar Dwivedi explored computational intelligence techniques for predicting diabetes mellitus in an article titled "Analysis of Computational Intelligence Techniques for Diabetes Mellitus Prediction," published in Neural Computing Applications in 2017.

[6] Bo Jin, Chao Che, Zhen Liu, Shillong Zhang, Xiaomeng Yin, And Xiaoping Wii, "Predicting the Risk of Heart Failure With EHR Sequential Data Modelling", IEEE Access 2018.

[7] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", IEEE Access 2019.

[8] A. Lakshmanarao, Y. Swathi, and P. Sri Sai Sundareswar investigated the application of machine learning techniques in predicting heart disease. This research was published in the International Journal of Scientific & Technology Research, Volume 8, Issue 11, in November 2019.

[9] Mijwil, M. M., Al-Mistarehi, A. H., and Mutur, D. S. conducted a literature review titled "The Practices of Artificial Intelligence Techniques and Their Value in Addressing the COVID-19 Pandemic" in the Mobile Forensics Journal in 2022, volume 4, issue 1, pages 11-30.

[10] A. Taneja authored a paper titled "Heart Disease Prediction System Using Data Mining Techniques" in the Oriental Journal in 2013.

[11] N. O. Fowler published a paper on the diagnosis of heart disease in volume V of an unspecified journal in March 2012.

[12] Ashok Kumar Dwivedi explored computational intelligence techniques for predicting diabetes mellitus in an article titled "Analysis of Computational Intelligence Techniques for Diabetes Mellitus Prediction," published in Neural Computing Applications in 2017.

[13] M. Shahi and R. Kaur Germ presented a heart disease prediction system using data mining techniques in the Orient Journal of Computer Science and Technology in 2013.

[14] S. M. S. Shah et al. conducted research on heart disease diagnosis using parallel probabilistic principal component analysis to extract relevant features, published in Statistical Mechanics and its Applications in 2017.

[15] T. Karthikeyan, B. Raghavan, and V. A. Kanimozhi conducted a study on the application of data mining classification algorithms for heart disease prediction, which was published in the International Journal of Advanced Research in Computer Engineering and Technology, volume 5, issue 4, pages 1076–1081, also in 2016.

[16] Folsom, A. R., Princes, R. J., Kaye, S. A., and Solar, J. T. (1989) investigated the association between body fat distribution and the self-reported prevalence of hypertension, heart attack, and other heart diseases in older women.

[17] Gour, S., Panwar, P., Dwivedi, D., and Mali, C. (2022). In 'Intelligent Sustainable Systems,' published by Springer in Singapore, the chapter titled 'A Machine Learning Approach for Heart Attack Prediction' spans pages 741 to 747."

[18] Konda Babu, A., Siddhartha, V., Kumar, B.B., and Penumutchi, B. (2021). "A comparative study on machine learning-based heart disease prediction." Materials Today Proceedings.

[19] Gudmundsson, E.F., Bjornsdottir, G., Sigurdsson, S., Andersen, K., Thorsson, B., Aspelund, T., and Gudnason, V. (2022). In a population-based cohort, the presence of carotid plaque correlates significantly with coronary artery calcium and serves as a predictive factor for the development of incident coronary heart disease.

[19] A. Jagtap, P. Malewadkar, O. Baswat, and H. Rambade conducted a study titled "Heart disease prediction using machine learning," published in the International Journal of Research in Engineering, Science, and Management in 2019.

[20] U. N. Dulhare published a paper titled "Prediction system for heart disease using naive Bayes and particle swarm optimization" in the Biomedical Research Journal in 2018.

[21] J. K. Kim and S. Kang presented research on "Neural network-based coronary heart disease risk prediction using feature correlation analysis" in the Journal of Healthcare Engineering in 2017.

[22] K. C. Siontis, P. A. Noseworthy, Z. I. Attia, and A. Paul discussed "Artificial intelligence-enhanced electrocardiography in cardiovascular disease management" in a 2021 article published in Nature Reviews Cardiology.

[23] P. S. Linda, W. Yin, P. A. Gregory, Z. Amanda, and G. Margaux developed "Development of a Novel Clinical Decision Support System for Exercise Prescription among Patients with multiple cardiovascular disease risk factors," published in Mayo Clinic Proceedings: Innovations, Quality & Outcomes in 2021.

[24] Y. Ali, R. Amir, and A.-M. Fardin conducted a study titled "Profile-based assessment of diseases affecting factors using fuzzy association rule mining approach: a case study in heart diseases," published in the Journal of Biomedical Informatics in 2021.

[25] Ghwanmeh, S., Mohammad, A., & Al-Ibrahim, A. (2013). "Innovative artificial neural network-based decision support system for heart disease diagnosis." Journal of Intelligent Learning Systems and Applications, 5(3), 353-396.

[26] Al-Shayea, Q. K. (2011). "Artificial neural networks in medical diagnosis." International Journal of Computer Science Issues, 8(2), 150–154.