Comparative Analysis of Machine Learning Models for Predicting Heart Disease

[1] S Gowri Krishna,https://orcid.org/0009-0008-0343-0643
[2]Ajitha R. Subhamathi
[1]Nirmala College Muvattupuzha, Kerala, India
[2] NSS College Rajakumari, Idukki, Kerala, India

Corresponding Author: *S Gowri Krishna, (email: gowrirajakumary@gmail.com)

***Abstract*— Heart disease is currently one among the principal causes of death worldwide, with multiple factors contributing to its development, such as genetic predisposition, health conditions, and lifestyle choices. When the narrowing of a heart valve exceeds 50%, the patient is typically considered to be at risk for cardiac arrest. Detecting early signs that could lead to heart attacks and related conditions has become a key focus of research. This study compares the performance of various classification models in predicting heart disease, including random forests, logistic regression, support vector machines, and decision trees. The results from this experimental analysis indicate that logistic regression shows higher precision, while random forests achieve higher accuracy compared to the other models. Random forests also demonstrate better functioning in terms of both accuracy and recall. Both models outperform the others when evaluated by F1 Score. Hyperparameters are optimized using grid search.**

***Index Terms*— heart disease prediction; machine Learning**

I. INTRODUCTION

Heart disease, also implied to as cardiovascular disease, comprises a broad range of conditions that influence the circulatory system, including the blood vessels and the heart. It is one of the most major sources of disability and mortality worldwide. Conditions such as blocked or narrowed coronary arteries, malfunctions of heart valves, and the enlargement of the heart muscle are key contributors to heart failure and heart attacks. As these conditions progress, they weaken the heart's ability to function properly, leading to life-threatening complications.

In recent years, machine learning (ML) has come up as a powerful tool in the field of disease prediction, coming up with new ways to improve accuracy. By selecting and analyzing certain features of cardiovascular health, ML models can enhance relevance of predictions. However, one of the challenges lies in the early detection of heart disease. Detecting the likelihood of heart disease at its earliest stages enables timely interventions, reducing the risk of progression to life-threatening conditions. Hence, developing models for early diagnosis has become a major focus in both the medical and machine learning communities.

In [1], a novel approach is introduced: preprocessing data to remove noise and optimize inputs,also hyperparameter optimization to fine-tune the model for better prediction results. The study in [2] uses k-modes clustering, enhanced by Huang's initialization method, to improve the accuracy of classification tasks. This method allows for more precise grouping of patients based on shared characteristics.

Additionally, several models, such as XGBoost (XGB), have been examined to identify key variables that are critical for predicting heart disease. In [23], a stacking model is presented, which includes multiple machine learning algorithms, including deep learning techniques, which increases the robustness and reliability of predictions. The model undergoes K-Fold cross-validation, which ensures that it is tested on multiple subsets of data, leading to more generalizable and accurate results. This stacking model demonstrates an improvement in predictive performance, all while maintaining complexity and accuracy.

Further advancements are discussed in [4], where a hybrid approach is used to detect cardiovascular diseases. By integrating different machine learning algorithms, this hybrid model can advance the strengths of each technique, finally providing a more comprehensive prediction system for heart disease.

Summary of the most recent research efforts is provided in [20]. This study explains how clustering algorithms and the application of various hyperparameter tuning techniques can substantially perfect the accuracy of heart disease prediction systems across a selection of clinical scenarios. By fine-tuning the model parameters, researchers can optimize model performance for different patient datasets and medical conditions, leading to more reliable predictions.

In [5], a deep learning-based artificial neural network (ANN) model is used to predict heart disease. Deep learning models like ANN have shown remarkable potential in processing large datasets and complex patterns that are often missed by traditional algorithms. The study demonstrates how deep learning techniques operate effective than conventional machine learning models in terms of accuracy, especially when large amounts of clinical data are available for training the models.

Overall, the comparison between traditional machine learning models and deep learning algorithms reveals that while conventional models show satisfactory results, the deep learning algorithms consistently have better results, particularly in terms of prediction accuracy. This study compares the effectiveness of different machine learning classification models for heart disease prediction, emphasizing the

performance improvements that can be achieved through careful hyperparameter optimization and model selection. The key findings and results of this study are detailed in Section IV, where the performance metrics of the different models are summarized and analyzed.

This expanded explanation provides a clearer view of the ongoing research in heart disease prediction and the promise of machine learning to transform early diagnosis and intervention in cardiovascular healthcare.

II. BASIC CONCEPTS

This study is to assess the performance of different machine learning algorithms, which are logistic regression, random forests, decision trees, and support vector machines, in predicting heart disease. The section provides an overview of the machine learning models used and discusses their outcomes.

Logistic Regression:

This supervised learning algorithm is generally used for binary classification problems. The limitation observed in this model is that it performs better only when the number of target variables is small.

Naive Bayes:

This algorithm is based on Bayes' theorem that predicts outcome based on the probability of events.

Decision Tree:

A decision tree is a broadly used and powerful algorithm for both classification and regression. It is a supervised learning method that visualizes decisions and their possible consequences through a tree-like structure.

Random Forest:

Random forest constructs numerous decision trees during the training phase, with each tree constructed using different variables and classifications suited for handling categorical data and multiclass problems, as the result is from the analysis and comparison of all the decision trees.

Support Vector Machine (SVM):

SVM is a classification algorithm like logistic regression. The algorithm uses hyperplanes to divide the dataset and is particularly effective with multidimensional data. The algorithm is more effective for complex datasets.

III. MATERIALS AND METHODS

A. Dataset Description

This study utilizes data related to heart disease to associate the accuracy of various machine learning models in predicting heart conditions. The necessary data has been sourced from the UCI dataset available on Kaggle. This dataset contains key features that are crucial for detecting heart attacks and narrowing blood vessels. These comprehensive features serve as a valuable resource for developing heart disease prediction models and conducting detailed analyses. A summary of the dataset's features is provided in Table 1.

Table 1. Heart disease dataset description

Feature	Description
Age	Age in years
sex	Sex (0 = female, 1 = male)
cp	Type of chest pain
trestbps	Blood pressure at rest (measured in mm Hg when the patient is admitted)
chol	Concentration of cholesterol in blood (mg/dl)
fbs	Blood glucose level in fasting (0 = false, 1 = true)
restecg	Results of resting ECG
thalach	Peak heart rate
exang	Angina triggered by exercise (0 = no, 1 = yes)
oldpeak	ST depression provoked by exercise relative to rest
slope	The slope of the peak exercise ST segment
ca	The number of major vessels (0-3) colored by fluo-roscopy.
thal	Thalassemia (0 = error, 1 = fixed defect, 2 = normal, 3 = reversible defect)
target	Disease label (0 = no disease, 1 = disease)

B. Evaluation Metrics

The following is a description of the metrics for evaluating machine learning models for prediction.

Confusion Matrix: This is used to illustrate the performance of a classification model.

Table 2. Confusion matrix

		Predicted	
		Negative	**Positive**
Actual	**Negative**	TN	FP
	Positive	FN	TP

Accuracy:

Accuracy is an ideal evaluation metric to be evaluated which is calculated as the ratio of correctly predicted ones to the total number.

Precision:

It is useful in real world situations where false positives are crucial. It provides the ratio of correctly predicted positive values to the total predicted positive ones.

Recall:

Also known as sensitivity, which calculates the ratio of correctly predicted positive values to the actual positives.

F1 Score:

The F1 score is the harmonic mean of precision and recall.

C. Covariance Matrix

The covariance matrix is a crucial tool in both statistics and machine learning, used to examine the relationships between variables. It is also referred to as the dispersion matrix or the variance-covariance matrix. Formally, if X is a dataset with nnn observations and ppp variables, the covariance matrix Σ\SigmaΣ is a p×pp \times pp×p matrix where each element Σij\Sigma_{ij}Σij represents the covariance between the iii-th and jjj-th variables.

In this expression, X′ and Y″ are to denote the mean value of X and Y, respectively. For this analysis, the Heart Disease UCI dataset from Kaggle is being used, and a detailed description of the dataset can be found in Section III A. As part of the data preprocessing, irrelevant information is pruned, followed by splitting the dataset into training and testing subsets. The machine learning models applied to this dataset include Random Forest, Logistic Regression, Naive Bayes, Decision Tree, and Support Vector Machine, each evaluated for their effectiveness in predicting heart disease. The covariance matrix is also analyzed to assess opportunities for dimensionality reduction. Hyperparameter tuning is conducted using both Grid Search Cross-Validation (GridSearchCV) and Randomized Search Cross-Validation (RandomizedSearchCV).

Table 3. Performance Comparison of Classification Models

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.81	0.78	0.89	0.83
Naive Bayes	0.78	0.75	0.85	0.80
Decision Tree	0.75	0.75	0.77	0.75
Random Forest	0.85	0.75	0.92	0.83
SVM	0.79	0.76	0.87	0.81

Performance comparison after Hyperparameter optimisation				
Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.86	0.88	0.89	0.85
Naive Bayes	0.87	0.90	0.85	0.87
Decision Tree	0.82	0.88	0.90	0.87
Random Forest	0.84	0.87	0.92	0.84
SVM	0.87	0.84	0.90	0.87

IV. RESULTS ANALYSIS

Random Forest, Logistic Regression, Decision Tree, Naive Bayes, and Support Vector Machine models are employed for predicting heart disease. The performance of these classification algorithms is evaluated with four metrics: accuracy, precision, recall, and F1 Score. The findings are presented in Table 3.

Logistic regression models are widely used for prediction as they effectively model the relationship between input features and a binary outcome. It learns a linear relationship from labeled data and the logistic function is used as an activation function to transform real-valued predictions into probabilities. In this experimental analysis Logistic regression gives better results in terms of precision. Random Forest achieves higher accuracy compared to the other models, as it combines predictions from several decision trees that are trained on random data subsets. This approach enhances generalization by mitigating overfitting. Additionally, it yields superior results in both Accuracy and Recall. Both models excel in terms of F1 Score when compared to the others.

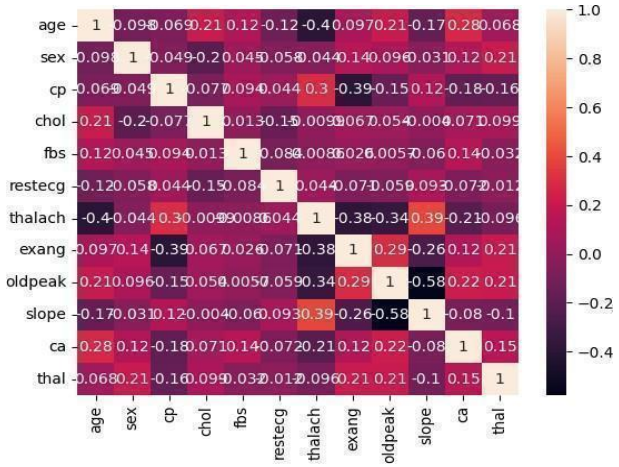

Fig.1. Covariance Matrix

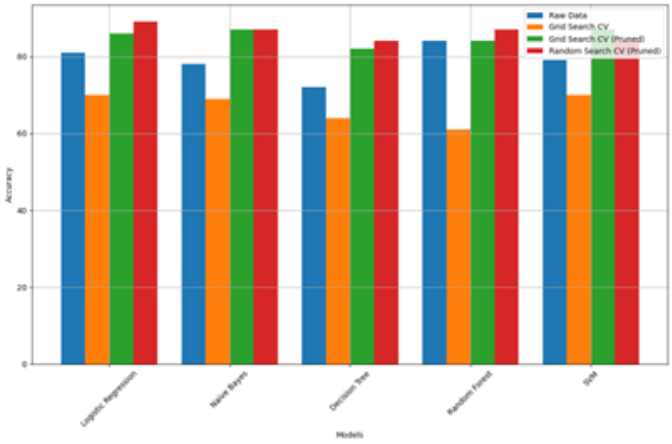

Fig.2. Accuracy after Hyperparameter Optimisation

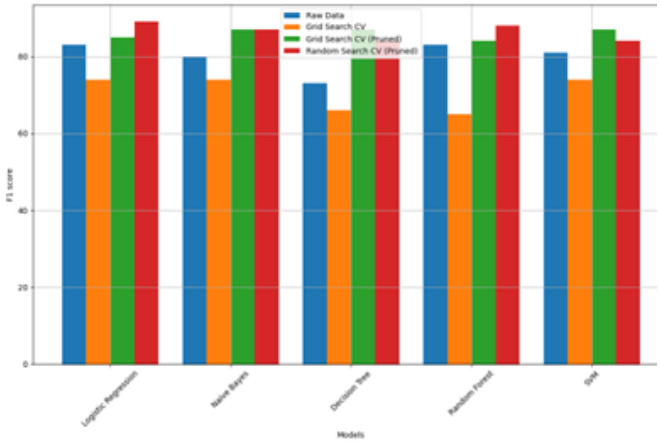

Fig.3. F1 Score After Hyperparameter Optimisation

Next, we look for the scope of dimensionality reduction techniques for better results. The covariance between each pair of variables is computed and is plotted in Fig.1. It can be observed from the figure that almost all values are in the range (-1,0), this proves that there is not much correlation between the variables. This poor correlation between variables suggests that dimensionality reduction techniques may not yield significant improvements in the performance of prediction models. Even though methods such as Principal Component Analysis result in dimensionality reduction and improve computational efficiency, they may not effectively capture relationships in datasets where variables are weakly correlated. Furthermore, the selection of relevant variables is very crucial when correlations are minimal, even advanced techniques like correlation networks may fail to identify informative features, which may lead to reducing the potential benefits of these dimensionality reduction techniques in improving model outcomes. So, we won't find any better result even if we use dimensionality reduction techniques.

Hyperparameter tuning significantly affects the performance of various prediction models. The performance of different prediction models in terms of accuracy and F1-Score after hyperparameter optimization is plotted in Fig.2. and Fig.3. respectively. The methods used for hypertuning were Grid searchCV and random searchCV are effective for different prediction models. GridSearchCV is time consuming as it systematically explores all possible combinations of hyperparameters. On the other hand, RandomizedSearchCV results in greater time efficiency since it randomly samples hyperparameter combinations. The classification models are trained for different iterations and the results were also analysed. It is observed that the model's performance is better when the number of iterations is fifty than when it was trained with less, or greater, nearing two hundred iterations. The selection of fifty iterations eliminates the possibility of overfitting and computational expenses.

V. CONCLUSION

In this study performance of different clustering algorithms were compared for heart disease. A better balance between Precision and recall for Random Forest Observed. It was observed that logistic regression shows better results in terms of precision. The methods used for hypertuning, Grid searchCV and random searchCV, are effective for different prediction models. The number of iterations was altered, and the results were analyzed, it is observed that when the number of iterations was too low or too high, the results were not good. A moderate number of iterations eliminates the possibility of overfitting and computational expenses.

REFERENCES

[1] Daniyal Asif,Mairaj Bibi,Muhammad Arif,Aiman Mukheimer, "Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization", Algorithms, 2023', 10.3390/a16060308

[2] Chintan Bhatt,Parth Patel,Tarang Ghetia,Pier Luigi Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques", Algorithms, 2023', 10.3390/a16020088

[3] Khader Basha Sk,D Roja,Sunkara Santhi Priya,Lavanya Dalavi,Sai Srinivas Vellela,Venkateswara Reddy B, "Coronary Heart Disease Prediction and Classification using Hybrid Machine Learning Algorithms", 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), 2023'.

[4] Ashish Singer and KK Hiran-Hybrid Approach for Heart Disease Detec- tion using classification Algorithms-2023 IEEE International Conference on ICT in Business Industry and Government

[5] Sarra, Raniya Rone, Dinar, Ahmed Musa, Mohammed, Mazin Abed, "Enhanced accuracy for heart disease prediction using artificial neural network", CERN European Organization for Nuclear Research - Zenodo, 2023', 10.11591/ijeecs.v29.i1. pp375-383

[6] Nadikatla Chandrasekhar,Samineni Peddakrishna, "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization", Processes, 2023', 10.3390/pr11041210

[7] Deepak Kumar,SK. Abdul Kareem,K. Roopesh,P. Aneesh, "Heart Disease Prediction Model", International Journal for Research in Applied Science and Engineering Technology, 2023', 10.22214/ijraset.2023.49570

[8] Mohamed G. El-Shafiey, Mohamed G. El-Shafiey, Ahmed Hagag, A. ElâDahshan,Manal A. Mahamod Ismail, "A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest", Multimedia Tools and Applications, 2022', 10.1007/s11042-022-12425-x

[9] Muhammad Salman Pathan, Avishek Nag, Muhammad Mohisn Pathan, Soumyabrata Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction", Healthcare analytics, 2022', 10.1016/j.health.2022.100060

[10] Huru Hasanova, Muhammad Tufail,, Ui-Jun Baek, Jee-Tae Park, Myung-Sup Kim, "A novel blockchain-enabled heart disease prediction mechanism using machine learning", Computers & Electrical Engineering, 2022', 10.1016/j. compeleceng. 2022.108086

[11] Raniya R. Sarra, Ahmed M. Dinar, Mazin Abed Mohammed, Karrar Hameed Abdulkareem, "Enhanced Heart Disease Prediction Based on Machine Learning and Ï‡2 Statistical Optimal Feature Selection Model", Designs, 2022', 10.3390/designs6050087

[12] P. Dileep, K. Nageswara Rao, Prajna Bodapati, Sitaratnam Gokuruboyina, Revathy Peddi,,Amit Grover,Amit Grover,Anu Sheetal,Anu Sheetal, "An automatic heart disease prediction using cluster-based bi-directional LSTM (C-BiLSTM) algorithm", Neural Computing and Applications, 2022', 10.1007/s00521-022-07064-0

[13] Ibrahim M. El-Hasnony,Ibrahim M. Elâ€Hasnony,Omar M. Elzeki,Omar M. Elzeki,Ali Alshehri,Ali Alshehri,Hanaa Salem,Hanaa Salem, "Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction", Sensors, 2022', 10.3390/s22031184

[14] Ali Al Bataineh,Ali Al Bataineh,Sarah Manacek,Sarah Manacek, "MLP-PSO Hybrid Algorithm for Heart Disease Prediction", Journal of Personalized Medicine, 2022', 10.3390/ jpm12081208.

[15] K. S. Ubale, P. N. Kalavadekar, "Effective Heart Disease Prediction Using Machine Learning Techniques", , 2021',

[16] Kaushalya Dissanayake,Kaushalya Dissanayake,Gapar Md Johar,Gapar Md Johar, "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms", Applied Computational Intelligence and Soft Computing, 2021', 10.1155/2021/5581806

[17] M. Kavitha,M. Kavitha,G. Gnaneswar,G. Gnaneswar, R. Dinesh,R. Dinesh,Yaozhang Sai,Y. Rohith Sai,R. Sai Suraj,R. Sai Suraj, "Heart Disease Prediction using Hybrid machine Learning Model", International Congress on Information and Communication Technology, 2021', 10.1109/ icict50816.2021 .9358597

[18] Sudarshan Nandy,Sudarshan Nandy,Sudarshan Nandy,Mainak Adhikari,Mainak Adhikari,Venki Balasubramanian,Venki Balasubramanian,Varun G. Menon,Varun G. Menon,Xingwang Li,Xingwang Li,Muhammad Zakarya,Muhammad Zakarya, "An intelligent heart disease prediction system based on swarm-artificial neural network", Neural Computing and Applications, 2021', 10.1007/s00521-021-06124-1

[19] Pooja Rani,Pooja Rani,Rajneesh Kumar Gujral,Rajneesh Kumar,Rajesh Kumar,Rajneesh Kumar,Nada Ahmed,Nada M. O. Sid Ahmed,Anurag Jain,Anurag Jain, "A decision support system for heart disease prediction based upon machine learning", Journal of Reliable Intelligent Environments, 2021', 10.1007/s40860-021-00133-6

[20] Abhay Agrahary,Abhay Agrahary, "Heart Disease Prediction Using Machine Learning Algorithms", International Journal of Scientific Research in Computer Science Engineering and Information Technology, 2020', 10.32628/cseit206421

[21] Jian Ping Li , Amin Ul Haq , Salah Ud Din , Jalaluddin Khan ,Asif Khan ,And Abdus Saboor -Heart Disease Identi- fication Method UsingMachine Learning Classification in E-Healthcare-DOI 10.1109/ACCESS.2020.3001149 , volume 8

[22] Bahzad Taha Jijo and Adnan Mohsin Abdulazeez-Classification Based on Decision Tree Algorithm forMachine Learning- Journal of Applied Science and Technology vol-2

[23] Mrs. B. Lalitha Rajeswari -Heart Disease Detection Using Machine Learning and Deep Learning-IJFANS International Journal Of Food And Nutritional Sciences.

[24] Montu, Saxena Tarun, Kaithwas Sanjana, Yadav Rahul and Lal Nidhi, Retracted: Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Ma- chine Learning 2020 International Conference on Computer Communication and Informatics (ICCCI), 2020, pages 1-6, doi=10.1109/ICCCI48352.2020.9104210