International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Enhanced Heart Disease Risk Prediction with Hyperparameter-Tuned Ensemble Models

Alok Singh Chauhan*[1], Mega Novita[2], Bambang Agus Herlambang[3], Dhyanendra Jain[4], Muhammad Saifuddin Zuhri[5], Priya Mishra[6], Dian Ayu Zahraini[7], Shikha Verma[8]

Submitted: 10/03/2024 **Revised:** 25/04/2024 **Accepted:** 02/05/2024

Abstract: Due to a number of risk factors, heart disease is a serious worldwide health concern that needs quick access to reliable early diagnosis and management techniques. Accurate prediction presents challenges, as seen in the limitations of traditional diagnostic methods. With the growing population, early-stage diagnosis becomes critical. Recent technological advancements have led to research in machine learning applications in healthcare, addressing these challenges. By examining pertinent variables, this work seeks to create an efficient machine learning model for the prediction of heart disease. A number of supervised learning techniques are used, such as XGBoost, K-Nearest Neighbor, Gradient Boosting, Random Forest, Decision Tree, and Logistic Regression. The primary goal is to estimate individuals' heart disease probability based on these factors. In this research, we overcome traditional diagnostic method limitations by utilizing ensemble methods, including the Gradient Boosting algorithm. This approach enhances heart disease prediction accuracy by integrating weak models. These methods open new avenues for heart disease management through detailed data analysis. The results show an impressive overall accuracy score of 99.02%. The developed model provides valuable insights, aiding informed decisions in diagnosis and treatment. Its integration into clinics supports early detection, potentially improving patient outcomes and reducing heart disease-related mortality. Beyond predictions, this study streamlines medical decision-making and revolutionizes heart disease care, enhancing patients' quality of life.

Keywords: *Heart disease prediction; Ensemble learning; Machine learning algorithms; Early diagnosis; Clinical decision-making*

1. Introduction

Heart disease stands as a pressing global health concern, remaining the leading cause of mortality worldwide [1–4]. Swift access to reliable, practical, and precise methods for early diagnosis and effective management is imperative due to the myriad of risk factors associated with heart disease. Despite strides in medical science, accurately predicting and diagnosing heart problems remains challenging, leading to delays in timely intervention and suboptimal patient outcomes. The intricate nature of heart disease necessitates a comprehensive understanding of various contributing factors. Traditional diagnostic methods, such as symptom-based assessments and standard medical tests, demonstrate limitations in accurately identifying heart disease due to subjective symptom interpretation, inter-observer variability, and an inability to capture subtle early signs. Hence, there is a critical need for advanced and accurate approaches to improve early detection, prognosis, and treatment decisions.

In recent years, technological advancements, particularly in artificial intelligence (AI) [5–8], have spurred research in healthcare. Within AI, machine learning (ML) has become a powerful technology that allows software applications to improve prediction accuracy without explicit programming. ML techniques enable the analysis of extensive datasets to extract valuable insights and make accurate predictions [9,10]. Heart disease prediction and treatment could be revolutionized by these methods, which include supervised, unsupervised, and ensemble learning classifiers. They have demonstrated promise in a number of domains. In order to predict future output values, machine learning algorithms use historical data. These algorithms are becoming essential to the healthcare industry since they enable accurate disease diagnosis and detection.

Machine learning algorithms have been used in several researches to forecast cardiac illness. Among the methods used were Logistic Regression [11], Random Forest [12], Support Vector Machine [13], K-Nearest Neighbors [14–16], and Gradient Boosting [17]. Of these, Logistic Regression proved to be the most successful, with an accuracy rate of almost 95%. Similarly, accuracy rates ranging from 80% to 98.2% have been attained in research employing several categorization methods, including Naive

[1]*School of Computer Applications and Technology, Galgotias University, Greater Noida, India*
[2]*Postgraduate Program of Science Education, Universitas PGRI Semarang, Indonesia*
[3]*Faculty of Engineering and Informatics, Universitas PGRI Semarang, Indonesia*
[4]*Department of CSE (AIML), ABES Engineering College, Ghaziabad, India*
[5]*Faculty of Mathematics, Universitas PGRI Semarang, Indonesia*
[6,8]*Department of Computer Applications, ABES Engineering College, Ghaziabad, India*
[7]*Faculty of Social Sciences and Sports Education, Universitas PGRI Semarang, Indonesia*
* *Corresponding Author Email: alok.chauhan@galgotiasuniversity.edu.in*

Bayes, Decision Trees, and Neural Networks. These studies [18–22] show how machine learning can improve the diagnosis and prognosis of heart diseases. However, there remains a gap in achieving even higher accuracy rates, particularly in the context of early detection. Additionally, integrating these techniques with specific medical expertise and understanding the impact of various factors on heart disease development could further enhance the precision of predictions. The research gap lies in refining existing methods, exploring potential optimization avenues, and incorporating a comprehensive understanding of both medical and machine learning domains.

This research addresses the critical need for early detection of heart-related disorders, aiming to mitigate their impact through timely medical guidance and interventions. Preventive measures, such as avoiding risky substances like alcohol and tobacco, engaging in regular physical activity, and maintaining a balanced diet, play a significant role in reducing heart disease risk. Early identification of these diseases is crucial for managing their impact on heart health. Factors including age, blood sugar levels, blood pressure, cholesterol, and other relevant medical characteristics are important in the development of heart disease. This study employs supervised machine learning to predict whether a patient, based on 13 medical characteristics, is likely to develop heart disease. The goal is to improve the precision and effectiveness of heart disease prediction by utilizing a variety of machine learning algorithms, such as XGBoost, Random Forest, Decision Tree, Gradient Boosting, Logistic Regression, Support Vector Machine, and K-Nearest Neighbor. By combining these methods with medical knowledge, this research aims to transform cardiovascular health management, allowing for individualized treatment regimens and timely interventions. The objectives include optimizing predictive performance by improving the accuracy and reliability of heart disease prediction through hyperparameter tuning of ensemble machine learning models, which is crucial for effectively identifying individuals at risk of heart disease and leading to earlier intervention and better patient outcomes. Additionally, enhancing model interpretability is important for building trust in predictive models and facilitating clinical decision-making, achieved by developing techniques to interpret ensemble machine learning models, enabling healthcare professionals to understand the factors influencing heart disease prediction. Understanding the features driving predictions can help clinicians tailor interventions and treatments for individual patients.

2. Materials and Methods

This study analyzes different machine learning techniques using a methodical approach to create an effective heart disease prediction system. In order to accurately detect heart disease, a number of algorithms were selected, including Random Forest [12], Decision Tree [23], Gradient Boosting [17], Logistic Regression[11], Support Vector Machine [13], XGBoost [24], and K-Nearest Neighbor [14–16]. These algorithms offer practitioners and medical analysts valuable insights.

In this research, we overcome traditional diagnostic method limitations by utilizing ensemble method, including the Gradient Boosting algorithm. This approach enhances heart disease prediction accuracy by integrating weak models, opening new avenues for heart disease management through detailed data analysis. The process initiates with data gathering, where a dataset containing pertinent information related to heart disease is collected. The dataset is then pre-processed to extract the most important features and get it ready for analysis. To guarantee data quality and consistency, problems including missing values, data cleansing, and data normalization are handled during this phase. Following pre-processing, the dataset is divided into two sections: 30% of the dataset is used for testing, while the remaining 70% is used for training. The test dataset is used to assess the machine learning classifiers' performance, and the training dataset is used to train them. The pre-processed data is subjected to the classifiers, which include Random Forest, Decision Tree, Gradient Boosting, Logistic Regression, Support Vector Machine, XGBoost, and K-Nearest Neighbor. These classifiers predict the chance of heart disease based on 13 medical characteristics, such as age, sex, blood pressure, cholesterol, fasting blood sugar, and chest discomfort[25–28].

To evaluate their efficacy, performance criteria like accuracy, precision, recall, F1 score (a combination of precision and recall), and Receiver Operating Characteristic - Area under the Curve (ROC-AUC) score are used. Based on the results, the best-performing algorithm is selected as the optimal choice for predicting heart disease. In this technical exploration, machine learning algorithms function as predictive tools, utilizing historical data to forecast new output values. In order to forecast the risk of heart disease, the study carefully looks at 13 medical characteristics, including age, sex, blood pressure, cholesterol, fasting blood sugar, and chest discomfort. Performance measures that show the predictive power of the algorithms include accuracy, precision, recall, F1 score, and ROC-AUC score. This research makes major contributions to the world of healthcare by improving the accuracy and efficiency of heart disease prediction by choosing the most efficient algorithm. The proposed methodology includes data pre-processing, feature engineering, data normalization, ensemble model building, hyperparameter optimization, and model evaluation, ensuring robust and reliable predictions.

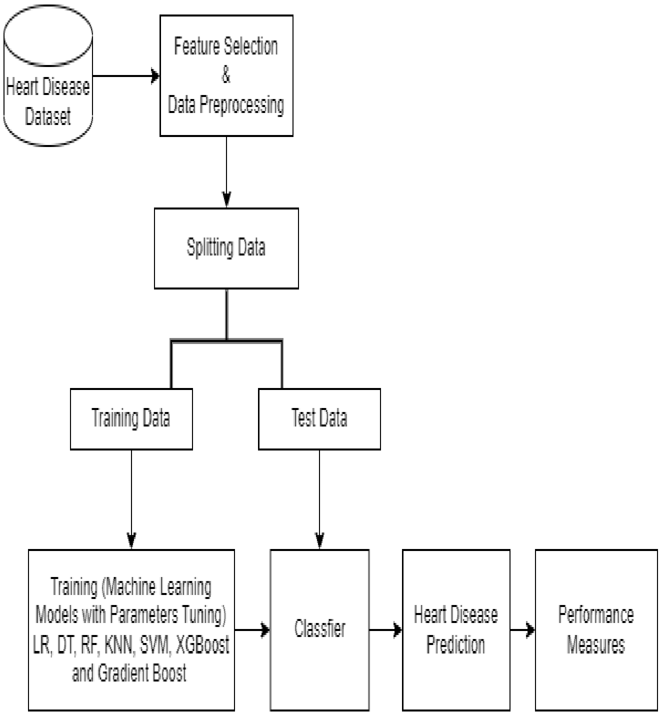

Figure 1. Research flow for developing the heart disease prediction system

3. Results and Discussion

Exploratory data analysis is part of the first stage of the research process to ensure reliable results. Prior to evaluating machine learning techniques, this investigation focuses on comprehending the features of the heart disease dataset. The patient's medical history and other pertinent data make up the dataset used in this study, which was downloaded from the Kaggle repository [29].

A thorough summary of the organization of the heart disease dataset included in this study may be found in Table 1. The dataset includes fourteen variables, or attributes, that record different facets of a patient's medical background as well as other pertinent information on heart disease. Every characteristic is useful in determining whether cardiac disease is present or not.

Below is a comprehensive explanation of every attribute included in Table 1.

1. age: The person's age in years is represented by this characteristic. Given that the chance of developing heart problems rises with age, age is a critical element in determining one's risk for heart disease.
2. sex: The patient's gender is indicated by this characteristic. The patient's sex is indicated by an encoded value of 0 for men and 1 for women. This information can be important for analyzing cardiac disease since risk factors may differ depending on the patient's gender.
3. CP (chest discomfort): This characteristic characterizes the patient's particular kind of chest discomfort. It can have one of four values—0, 1, 2, or 3—to represent various kinds and levels of discomfort or chest pain.
4. restbps (resting blood pressure): This characteristic shows the patient's blood pressure in millimeters of mercury (mm Hg) while they are at rest. Blood pressure at rest is a crucial sign of cardiovascular health.
5. chol (blood cholesterol level): This characteristic indicates the blood cholesterol level of the patient, expressed as mg/dL, or milligrams per deciliter. An higher risk of heart disease is linked to elevated cholesterol levels.
6. fasting blood sugar level (fbs): The patient's blood sugar level during fasting is indicated by this property. If the fasting blood sugar is more than 120 mg/dL, it is shown as 1, and if not, it is shown as 0. An important risk factor for heart disease is diabetes, which can be indicated by elevated fasting blood sugar.
7. resting electrocardiographic readings (restecg): This characteristic shows the ECG readings obtained while the patient is at rest. It can have one of three values: 0, 1, or 2, which represent various patterns of ECG findings at rest.
8. thalach (maximum heart rate achieved): This characteristic shows the patient's highest heart rate attained during a particular activity. It is a crucial sign of overall health and cardiovascular fitness.
9. exang (exercise-induced angina): This characteristic shows if the patient has exercise-induced angina, which is characterized by discomfort or pain in the chest. Yes is represented by 1 and no by 0.
10. oldpeak (compared to rest, ST depression generated by activity): This characteristic quantifies the relative amount of ST depression brought on by exercise. An important finding on an ECG is ST depression, which may point to heart-related issues.
11. slope (slope of the peak exercise ST segment): The peak workout's ST section's slope is represented by this characteristic. During exercise, it can have values of 0, 1, or 2, which represent various angles of the ST segment.
12. ca (number of major vessels colored by fluoroscopy): The number of major blood vessels (0–4) that have been fluoroscopically colored is indicated by this feature. It offers details about blood artery blockages, which are common in heart disease, including their degree and presence.
13. thal (thalassemia): This characteristic stands for the blood condition thalassemia. It can have three values: normal (value 3), permanent defect (value 6), or reversible defect (value 7-8). Thalassemia may have an impact on heart health as well as the blood's ability to deliver oxygen.
14. target: The target column is where the forecast is made. It is binary, with 1 denoting the existence of cardiac disease and 0 denoting its absence. Machine learning algorithms try to predict this property from the other features in the dataset.

Table 1.Structure of the Heart Disease Dataset

Attributes	Description
Age	Person's Age (in years)
sex	Patient's gender (0 = male, 1 = female)
cp	Type of chest discomfort (4 values: 0, 1, 2, 3)
trestbps	Blood pressure at rest (in mm Hg)
chol	Blood cholesterol level (in mg/dl)
fbs	Blood sugar level at fasting > 120 mg/dl (1 = true; 0 = false)
restecg	Electrocardiographic readings while at rest (values 0, 1, 2)
thalach	Reached maximum heart rate
exang	Angina brought on by exercise (1 = yes; 0 = no)
oldpeak	Exercise-induced ST depression compared to rest
slope	The angle of the ST section of the peak workout (values 0, 1, 2)
ca	main vessels (0–4) coloured by flourosopy
thal	(3=normal; 6 = permanent defect; 7–8 = reversible defect);
target	Target column (1 equals Yes, 0 equals No)

Additionally, Table 2 presents a subset of the Heart Disease dataset, offering a glimpse into the data with 5 sample observations. Each row represents a unique patient, and the columns include precise information about individual patients that is critical for forecasting the existence or absence of cardiac disease. The target property, which reflects the presence or absence of cardiac disease, has 1,025 observations. Among these observations, 499 are tagged as having no heart illness (0), whereas 526 are labeled as having heart disease (1).

Table 2. Heart disease dataset

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	taget
52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
58	0	0	100	248	0	0	122	0	1	1	0	2	1

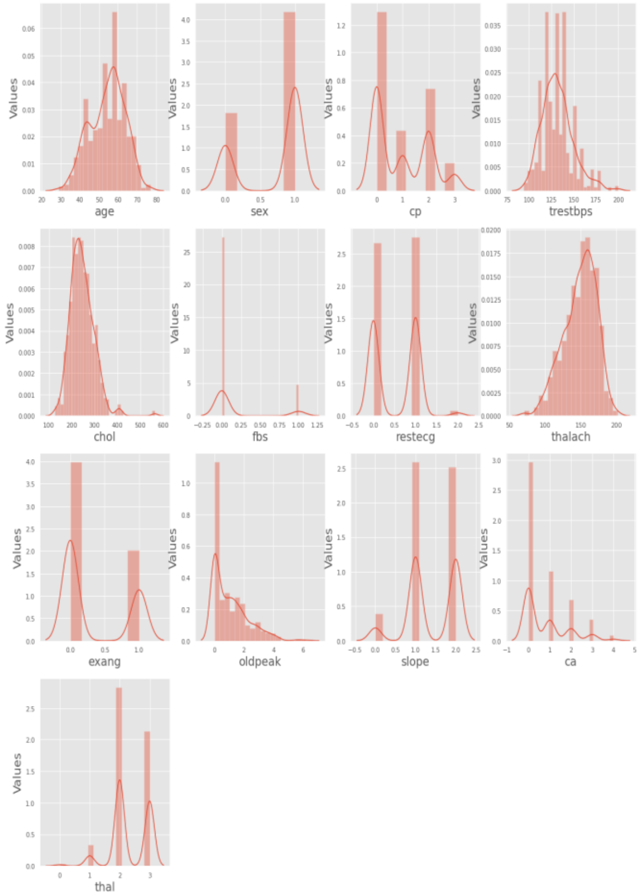

Figure 2. Data distribution across 13 variables such as age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, and thal

Figure 2 depicts the distribution of data across 13 variables important for predicting heart disease, allowing for a more comprehensive analysis of the data. These variables, including age, sex, chest discomfort (cp), resting blood pressure (trestbps), cholesterol levels (chol), fasting blood sugar (fbs), resting electrocardiographic readings (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), slope of the peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (ca), thalassemia type (thal), and the target variable. Understanding these patterns is critical to creating accurate machine learning models. Outliers or odd data points can have a substantial impact on model performance, and identifying them ahead of time allows for data cleaning and preprocessing, which ensures the prediction algorithms' reliability. Thus, Figure 2 is critical in preparing the data for robust and precise machine learning analysis.

Figure 3 presents a correlation matrix heat map, offering a visual depiction of the relationships between different variables in the dataset. This visualization illustrates both the strength and direction of correlations, aiding in the identification of potential dependencies and interrelationships among the variables. The heat map provides essential insights into the degree and nature of correlations, allowing researchers to discern patterns of positive and negative associations. A larger value inside the square suggests a better correlation between the variables it represents. Positive correlation shows that as one variable increases, the other variable tends to rise, whereas negative correlation indicates that as one variable rises, the other decreases. Researchers obtain a thorough understanding of the data by performing exploratory data analysis and visualizing its properties. This understanding is critical for making sound judgments about data pre-processing, feature selection, and model building. Analyzing the correlations allows researchers to fine-tune their models, ensuring that the chosen factors are meaningful and contribute considerably to the accuracy and effectiveness of the heart disease prediction system.

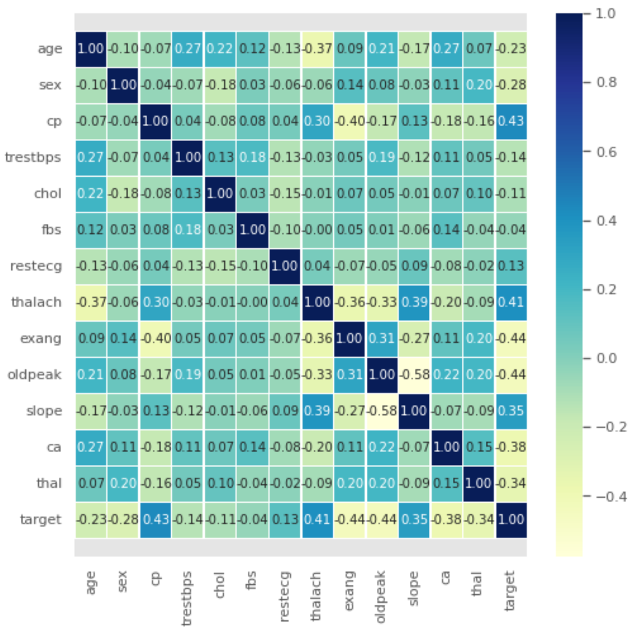

Figure 3. Correlation Matrix Heat Map

Following a thorough exploratory data analysis, seven machine learning algorithms (Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Decision Tree, Random Forest, Gradient Boosting, and XGBoost) were rigorously evaluated on a 70% training dataset and then tested on a 30% test dataset. The algorithms' performance was assessed using evaluation criteria such as accuracy score, precision, recall, F1 score, and area under the Receiver Operating Characteristic (ROC) curve.

The results, presented in Table 3, provide a full summary of the algorithms' performance. The Gradient Boosting Classifier demonstrated great accuracy, precision, recall, and F1 score, with an impressive accuracy rate of 99.02%, indicating its capacity to effectively diagnose heart disease. Furthermore, the classifier achieved a precision rate of 98.10%, demonstrating that it can reliably identify individuals with heart disease. The recall rate of 100% indicates that the classifier correctly identified all cases of heart disease, whereas the F1 score of 99.08% indicates that it performed well overall. Furthermore, the area under the ROC curve, which measures the classifier's ability to distinguish between positive and negative examples, was good, showing strong predictive power.

Comparing the different algorithms, Random Forest, Decision Tree, Gradient Boosting, and XGBoost consistently outperformed other classifiers. The Random Forest classifiers achieved a comparable accuracy rate of 98.70%, positioning them as the second-best performing algorithms overall. The Decision Tree Classifier demonstrated an accuracy rate of 97.72%, making it the third most accurate algorithm, and the XGBoost Classifier exhibited an accuracy rate of 97.40%, ranking it as the fourth most accurate among the evaluated models.

These findings align with previous studies emphasizing the effectiveness of Random Forest, Decision Tree, Gradient Boosting, and XGBoost in predicting heart disease [30–32]. Particularly, the Gradient Boosting Classifier emerged as the most accurate algorithm, surpassing accuracies reported in earlier research. Its 99.02% accuracy suggests its reliability in predicting heart disease presence in patients. The high precision, recall, and F1 score further underscore its robust performance. These findings show the ability of machine learning algorithms, specifically the Random Forest Classifier, to effectively predict cardiac disease. These prediction models considerably help medical practitioners diagnose and manage heart diseases, resulting in more effective therapies and better patient outcomes. The excellent accuracy rates achieved by these algorithms illustrate their capacity to correctly identify patients at risk of heart disease, allowing for early intervention and preventive actions .Figure 3's correlation insights probably helped with feature selection and optimization, which increased the predictive power of the algorithms. Table 3 illustrates how data visualization and exploratory analysis have a direct impact on the accuracy and performance of machine learning models, highlighting the importance of comprehending the nuances of the dataset.

Table 3. Performance metrics of different machine learning algorithms for heart disease prediction

Machine Learning Models	Accuracy	Precision	Recall	F1 Score	ROC-AUC Score
Logistic Regression (LR)	0.8831	0.8629	0.9264	0.8934	0.8804
K-Nearest Neighbor (KNN)	0.8603	0.8896	0.8405	0.8643	0.8969
Support Vector Machine (SVM)	0.8603	0.8297	0.9264	0.8753	0.8563
Decision Tree (DT)	0.9772	0.9875	0.9693	0.9783	0.9778
Random Forest (RF)	0.9870	0.9760	1.0000	0.9878	0.9862
Gradient Boosting (GBDT)	0.9902	0.9810	1.0000	0.9908	0.9897
XGBoost	0.9740	0.9874	0.9632	0.9751	0.9747

The ROC curves for the several predictive models used to forecast cardiac disease are shown in Figure 4. The trade-off between the true positive rate (sensitivity) and the false positive rate (specificity) for various categorization thresholds is represented graphically by the ROC curve. It offers important information about how well a model can discriminate between situations that are positive and cases that are negative over a variety of threshold values. The ROC curves in this particular context show how well various machine learning algorithms perform in terms of categorizing cases of heart disease. A greater Region The model's improved capacity to distinguish between those with heart disease (positive class) and those without (negative class) is indicated by the area under the ROC curve (AUC).

It is clear from examining Figure 4 that the AUC for the Random Forest (RF) and Gradient Boosting (GBDT) ROC curves is significantly higher than those of the other machine learning techniques. This shows that the GBDT and RF techniques are quite good at identifying the positive class in the dataset. Particularly, the GBDT and RF curves' AUC values show excellent sensitivity and specificity, indicating that these models can reliably identify people with and without heart disease.

The AUC statistic, which measures the classifier's capacity to differentiate between classes, is used to summarize the ROC curve study. Greater performance in differentiation is indicated by a higher AUC value. With the highest ROC-AUC score of 0.9897, the GBDT technique performed better in this instance than other models, demonstrating its remarkable ability to distinguish between cases of heart disease and non-heart illness. With a ROC-AUC score of 0.9862, the RF approach came in close second, demonstrating its superior classification efficacy.

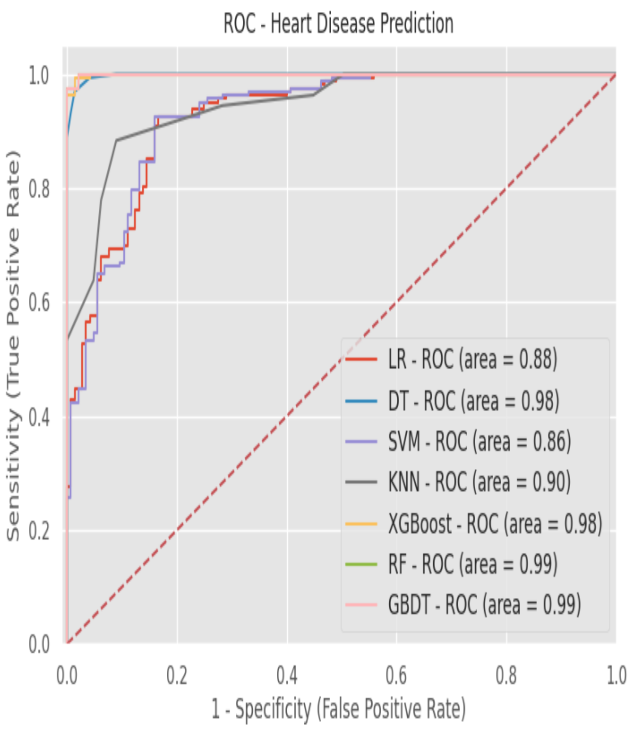

Figure 4. Receiver Operating Characteristic (ROC) curve for heart disease prediction

The Gradient Boosting (GBDT) algorithm's feature importance for heart disease prediction is shown in Figure 5. A machine learning statistic called feature importance is used to assess how important a given feature (or variable) is in producing precise predictions. This refers to the degree to which each feature affects the GBDT algorithm's capacity to forecast the target variable—heart disease, in this case—in this particular scenario.In this specific study, all 13 features, including fasting blood sugar (fbs), were considered. Despite fbs not traditionally being considered a significant feature, its inclusion in the analysis might have led to valuable insights. Sometimes, seemingly less significant features, when combined with other variables, can enhance the overall accuracy of the predictive model. Even though fbs isn't usually thought of as a major feature, having it in the study could have provided insightful information. When paired with additional variables, seemingly insignificant traits can occasionally improve the predictive model's overall accuracy.

This comprehensive approach, considering all available features, could have resulted in improved accuracy and a more holistic understanding of heart disease prediction. Chest pain (cp) emerges as the most influential factor in predicting heart disease according to the Gradient Boosting Algorithm. The severity and nature of chest pain experienced by an individual are crucial indicators. This feature likely carries significant weight in the model due to the well-established association between chest pain and heart-related issues. Patients reporting specific types of chest pain may have distinct underlying cardiac conditions, making this variable highly relevant for accurate predictions. In the prediction model, the number of main vessels colored by fluoroscopy (ca) is ranked second in importance. This variable indicates potential blockages or abnormalities in the major blood vessels supplying the heart. A higher count suggests a greater degree of vascular involvement, signifying a higher risk of heart disease. The more vessels affected, the more critical the condition, making this feature crucial in predicting heart-related ailments. The degree of exercise-induced ST depression compared to rest (oldpeak) and the type of thalassemia (thal) show almost comparable significance. Thalassemia is a genetic blood disorder that can affect the heart, altering its oxygen-carrying capacity. Similarly, ST depression during exercise (oldpeak) indicates abnormalities in the heart's electrical activity and is indicative of cardiac stress. The similarity in their importance underscores their complementary roles in assessing heart disease risk. These factors, when analyzed collectively, offer a comprehensive understanding of the patient's heart health. While chest pain serves as a direct symptom, the number of affected vessels, thalassemia type, and exercise-induced changes provide valuable insights into the underlying cardiovascular conditions. Even though they are important, other factors have a comparatively smaller impact on the model's predicted accuracy. Knowing the importance of these characteristics helps medical personnel diagnose and treat patients more precisely and quickly, protecting individuals at risk of heart disease.

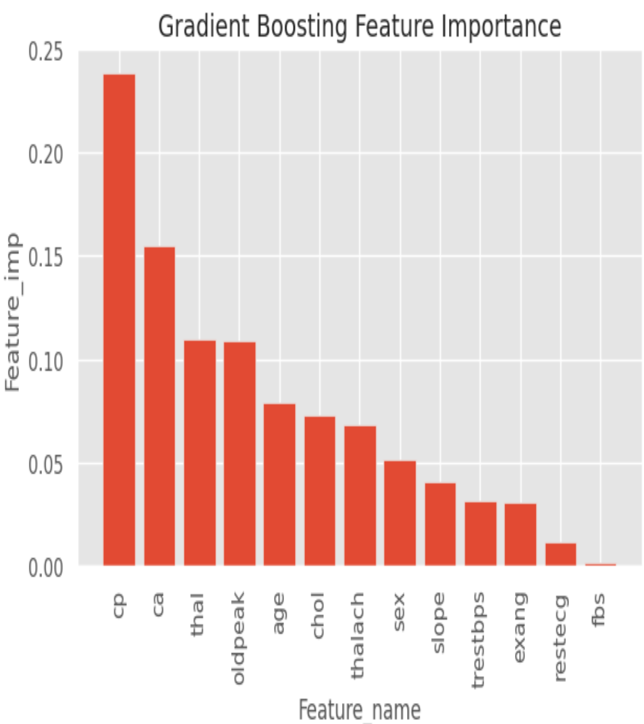

Figure 5. Feature Importance of Gradient Boosting Algorithm for heart disease prediction

This study's findings are noteworthy in a number of ways. First off, the Gradient Boosting Classifier's accuracy and precision, which exceed those of earlier studies, demonstrate how reliable it is in predicting the presence of heart disease. The influence of features like chest pain, vessel count, thalassemia type, and ST depression on predictions underscores their clinical relevance.

However, certain limitations need acknowledgment. Dataset limitations, potential biases, algorithm constraints, and temporal factors can impact the study's scope and generalizability [33–37]. Machine learning models, while powerful, require careful interpretation in clinical contexts, emphasizing the importance of expert oversight. Despite these limitations, the study's practical implications for healthcare settings are immense. Predictive models offer personalized, efficient, and proactive approaches to heart disease management. Integrating these models into healthcare applications can optimize patient care, enhance resource allocation, and empower medical professionals.

Additionally, the utilization of ensemble methods, particularly Gradient Boosting, significantly contributes to the predictive accuracy. Ensembles, combining predictions from multiple models, outperform individual models, offering robust and accurate results. The importance of understanding the dataset intricacies, as demonstrated through exploratory analysis, cannot be overstated. These discoveries improve the prognosis for cardiac disease and highlight the critical role that thorough data analysis and machine learning methods play in the medical field. To ensure continuing advancement in this crucial area, future research should concentrate on interpretability, incorporate sophisticated methodologies, and remove dataset biases.

To learn more about the characteristics and variables affecting the predictions, future studies in this area might investigate how interpretable the models are. Additionally, investigations incorporating more advanced techniques such as deep learning algorithms, alternative approaches for attribute selection, and larger datasets could further enhance the prediction accuracy and generalizability of the models.

The research findings hold significant practical implications for real-world healthcare settings, offering a range of benefits for both patients and medical professionals. Predictive models can serve as pivotal tools for early intervention and prevention strategies, enabling healthcare providers to identify individuals at high risk of heart disease even before symptoms manifest. Healthcare practitioners can identify patients at high risk of heart disease even before symptoms appear by using predictive models, which can be essential tools for early intervention and prevention strategies. By initiating timely interventions, such as lifestyle modifications or medication, the onset or progression of the condition can potentially be prevented. Moreover, these models facilitate the customization of treatment plans based on individual risk profiles [38–41]. This tailored approach ensures that interventions, whether dietary changes, exercise regimens, or specific medications, are precisely aligned with the patient's needs, optimizing their overall health outcomes.

In addition to personalized treatment strategies, predictive models enhance the efficient allocation of resources within healthcare facilities. Hospitals often face constraints, and these models help streamline the healthcare process by identifying patients who require immediate attention. By directing resources towards high-risk cases, the healthcare system operates more effectively, ensuring that critical resources are utilized where they are most needed. Furthermore, these models empower medical professionals by augmenting their diagnostic capabilities. By combining clinical expertise with machine-driven predictions, healthcare providers can make more informed decisions [42–45]. This synergy leads to more accurate diagnoses and effective treatments, ultimately improving patient care.

Practically, these models can be integrated into various healthcare applications [46–55]. For instance, personalized wellness apps can use these models to provide tailored suggestions for diet, exercise, and stress management based on users' heart disease risk profiles. In busy emergency departments, a predictive model integrated into the triage system assesses incoming patients' heart disease risk, ensuring that high-risk patients receive immediate attention. Additionally, remote monitoring devices equipped with predictive models can alert healthcare providers if a significant increase in heart disease risk is detected in patients with chronic conditions. This proactive approach facilitates timely interventions, even from a distance, and can potentially save lives in critical situations. By translating research findings into practical applications, these predictive models become invaluable tools in the hands of healthcare professionals, fostering a proactive, personalized, and efficient approach to heart disease management.

4. Conclusions

In light of the rising prevalence of cardiac issues, this study's findings emphasize the critical need of early identification. The main goal of the study was to accurately forecast heart disease by applying a variety of machine learning algorithms and conducting thorough data processing, with a focus on ensemble approaches. By harnessing clinical data pertaining to prior heart disease diagnoses, this method delivers substantial benefits to patients. Through the analysis of a dataset containing detailed patient medical histories, this approach adeptly identifies individuals at risk of developing potentially fatal heart conditions. Among the array of machine learning algorithms employed, Gradient Boosting emerged as the most powerful, boasting an exceptional accuracy rate of 99.02%. This reinforces its position as the most effective tool in predicting heart disease. Notably, the analysis revealed that the most influential factor in this prediction was 'cp' (chest pain type), signifying its pivotal role in identifying individuals susceptible to heart problems. The results of this investigation shed light on how well machine learning algorithms work when combined with different measures and models to accurately forecast heart disease. Future research endeavours could further refine prediction accuracy by integrating a multitude of algorithms and exploring additional influential factors. Predictive skills may also be strengthened by system improvements including the addition of Deep Learning algorithms, different methods for attribute selection, and a larger dataset.

Acknowledgments

This research was funded by Universitas PGRI Semarang grant number 007/SKK/LPPM-UPGRIS/KLN/I/2023.We extend our heartfelt gratitude to Universitas PGRI Semarang for providing the essential funding that made this research possible. This project stands as a testament to the collaborative spirit between Universitas PGRI Semarang, Indonesia, and ABES Engineering College, India. This joint effort exemplifies the power of international collaboration in advancing scientific knowledge and fostering academic exchange. We sincerely appreciate the support and encouragement extended to us throughout this research endeavor.

Author contributions

Alok Singh Chauhan 1: Conceptualization, Methodology, Software, Visualization, Investigation, Writing-Reviewing and Editing. **Mega Novita 2:** Data curation, Investigation, Writing-Original draft preparation. **Bambang Agus Herlambang 3:** Visualization, Investigation. **Dhanendra Jain 4**: Software, Visualization. **Muhammad Saifuddin Zuhri 5**: Field study. **Priya Mishra 6:** Field study, Writing-Reviewing and Editing. **Dian Ayu Zahraini 7:** Field study. **Shikha Verma 8**: Data curation, Writing-Reviewing and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Coffey S, Roberts-Thomson R, Brown A, Carapetis J, Chen M, Enriquez-Sarano M, Zühlke L and Prendergast B D 2021 Global epidemiology of valvular heart disease Nat. Rev. Cardiol.18

[2] Wu W, He J and Shao X 2020 Incidence and mortality trend of congenital heart disease at the global, regional, and national level, 1990-2017 Med. (United States)99

[3] Aluru J S, Barsouk A, Saginala K, Rawla P and Barsouk A 2022 Valvular Heart Disease Epidemiology Med. Sci. (Basel, Switzerland)10

[4] Khan M A, Hashim M J, Mustafa H, Baniyas M Y, Al Suwaidi S K B M, AlKatheeri R, Alblooshi F M K, Almatrooshi M E A H, Alzaabi M E H, Al Darmaki R S and Lootah S N A H 2020 Global Epidemiology of Ischemic Heart Disease: Results from the Global Burden of Disease Study Cureus

[5] Janiesch C, Zschech P and Heinrich K 2021 Machine learning and deep learning Electron. Mark.31

[6] Shukla A K, Janmaijaya M, Abraham A and ... 2019 Engineering applications of artificial intelligence: A bibliometric analysis of 30 years (1988–2018) … Appl. Artif. …

[7] Tran B X, Latkin C A, Vu G T, Nguyen H L T and ... 2019 The current research landscape of the application of artificial intelligence in managing cerebrovascular and heart diseases: A bibliometric and content analysis Int. J. …

[8] Guo Y, Hao Z, Zhao S, Gong J and Yang F 2020 Artificial intelligence in health care: bibliometric analysis J. Med. Internet …

[9] Jamwal A, Agrawal R, Sharma M, Kumar A and ... 2021 Machine learning applications for sustainable manufacturing: A bibliometric-based review for future research J. Enterp. …

[10] Stoll A and Benner P 2021 Machine learning for material characterization with an application for predicting mechanical properties GAMM Mitteilungen44

[11] Chatterjee S and Simonoff J S 2013 Handbook of Regression Analysis

[12] Cutler D R, Edwards T C, Beard K H, Cutler A, Hess K T, Gibson J and Lawler J J 2007 Random forests for classification in ecology Ecology88

[13] Alcaraz J, Labbé M and Landete M 2022 Support Vector Machine with feature selection: A multiobjective approach Expert Syst. Appl.204

[14] Maillo J, Garcia S, Luengo J, Herrera F and Triguero I 2020 Fast and Scalable Approaches to Accelerate the Fuzzy k-Nearest Neighbors Classifier for Big Data IEEE Trans. Fuzzy Syst.28

[15] Dinesh P and P. K 2022 Medical Image Prediction for Diagnosis of Breast Cancer Disease Comparing the Machine Learning Algorithms: SVM, KNN, Logistic Regression, Random Forest, and Decision Tree to Measure Accuracy ECS Trans.107

[16] Fadilah Z and Murnawan 2021 Performance Comparison of K-Nearest Neighbor and Decision Tree C4.5 by Utilizing Particle Swarm Optimization for Prediction of Liver Disease Int. J. Open Inf. Technol.9

[17] Feng X, Hua Y, Zou J, Jia S, Ji J, Xing Y, Zhou J and Liao J 2022 Intelligible Models for HealthCare: Predicting the Probability of 6-Month Unfavorable

Outcome in Patients with Ischemic Stroke Neuroinformatics20

[18] Mohan S, Thirumalai C and Srivastava G 2019 Effective heart disease prediction using hybrid machine learning techniques IEEE Access7

[19] Shorewala V 2021 Early detection of coronary heart disease using ensemble techniques Informatics Med. Unlocked26

[20] Ayano Y M, Schwenker F, Dufera B D and Debelee T G 2023 Interpretable Machine Learning Techniques in ECG-Based Heart Disease Classification: A Systematic Review Diagnostics13

[21] Ahsan M M and Siddique Z 2022 Machine learning-based heart disease diagnosis: A systematic literature review Artif. Intell. Med.128

[22] Fatima M and Pasha M 2017 Survey of Machine Learning Algorithms for Disease Diagnostic J. Intell. Learn. Syst. Appl.09

[23] Ingre B, Yadav A and Soni A K 2018 Decision tree based intrusion detection system for NSL-KDD dataset Smart Innovation, Systems and Technologies vol 84

[24] Budholiya K, Shrivastava S K and Sharma V 2022 An optimized XGBoost based diagnostic system for effective prediction of heart disease J. King Saud Univ. - Comput. Inf. Sci.34

[25] Fawcett T 2006 An introduction to ROC analysis Pattern Recognit. Lett.27

[26] Cort J. Willmott and Kenji Matsuura 2005 Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance Clim. Res.30

[27] Hodson T O 2022 Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not Geosci. Model Dev.15

[28] POWERS D M . 2011 EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION J. Mach. Learn. Technol.2

[29] Kaggle 2021 Kaggle: Your Home for Data Science Website

[30] Chang V, Bhavani V R, Xu A Q and Hossain M A 2022 An artificial intelligence model for heart disease detection using machine learning algorithms Healthc. Anal.2

[31] K. Pazhanivel, A. Pravin, M. Sakthiganesh and M. Saravanan 2023 Diagnosis of Heart Disease using Machine Learning Algorithms Int. J. Adv. Res. Sci. Commun. Technol.

[32] Ghumbre S U and Ghatol A A 2012 Heart disease diagnosis using machine learning Algorithm Advances in Intelligent and Soft Computing vol 132 AISC

[33] Dey D, Slomka P J, Leeson P, Comaniciu D, Shrestha S, Sengupta P P and Marwick T H 2019 Artificial Intelligence in Cardiovascular Imaging: JACC State-of-the-Art Review J. Am. Coll. Cardiol.73

[34] Johnson K W, Torres Soto J, Glicksberg B S, Shameer K, Miotto R, Ali M, Ashley E and Dudley J T 2018 Artificial Intelligence in Cardiology J. Am. Coll. Cardiol.71

[35] Attia Z I, Kapa S, Lopez-Jimenez F, McKie P M, Ladewig D J, Satam G, Pellikka P A, Enriquez-Sarano M, Noseworthy P A, Munger T M, Asirvatham S J, Scott C G, Carter R E and Friedman P A 2019 Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram Nat. Med.25

[36] Zhou B, Khosla A, Lapedriza A, Oliva A and Torralba A 2016 Learning Deep Features for Discriminative Localization Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition vol 2016-December

[37] Obermeyer Z and Emanuel E J 2016 Predicting the Future — Big Data, Machine Learning, and Clinical Medicine N. Engl. J. Med.375

[38] Poplin R, Varadarajan A V., Blumer K, Liu Y, McConnell M V., Corrado G S, Peng L and Webster D R 2018 Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning Nat. Biomed. Eng.2

[39] Esteva A, Kuprel B, Novoa R A, Ko J, Swetter S M, Blau H M and Thrun S 2017 Dermatologist-level classification of skin cancer with deep neural networks Nature542

[40] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S and Dean J 2019 A guide to deep learning in healthcare Nat. Med.25

[41] Min S, Lee B and Yoon S 2017 Deep learning in bioinformatics Brief. Bioinform.18

[42] Topol E J 2019 High-performance medicine: the convergence of human and artificial intelligence Nat. Med.25

[43] Rajpurkar P, Irvin J, Ball R L, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C P, Patel B N, Yeom K W, Shpanskaya K, Blankenberg F G, Seekins J, Amrhein T J, Mong D A, Halabi S S, Zucker E J, Ng A Y and Lungren M P 2018 Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists PLoS Med.15

[44] Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B and Yang G Z 2017 Deep Learning for Health Informatics IEEE J. Biomed. Heal. Informatics21

[45] Arbabshirani M R, Plis S, Sui J and Calhoun V D 2017 Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls Neuroimage145

[46] Liu X, Faes L, Kale A U, Wagner S K, Fu D J, Bruynseels A, Mahendiran T, Moraes G, Shamdas M, Kern C, Ledsam J R, Schmid M K, Balaskas K, Topol E J, Bachmann L M, Keane P A and Denniston A K 2019 A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis Lancet Digit. Heal.1

[47] McKinney S M, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado G C, Darzi A, Etemadi M, Garcia-Vicente F, Gilbert F J, Halling-Brown M, Hassabis D, Jansen S, Karthikesalingam A, Kelly C J, King D, Ledsam J R, Melnick D, Mostofi H, Peng L, Reicher J J, Romera-Paredes B, Sidebottom R, Suleyman M, Tse D, Young K C, De Fauw J and Shetty S 2020 International evaluation of an AI system for breast cancer screening Nature577

[48] Gulshan V, Peng L, Coram M, Stumpe M C, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson P C, Mega J L and Webster D R 2016 Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs JAMA - J. Am. Med. Assoc.316

[49] Beaulieu-Jones B K, Wu Z S, Williams C, Lee R, Bhavnani S P, Byrd J B and Greene C S 2019 Privacy-preserving generative deep neural networks support clinical data sharing Circ. Cardiovasc. Qual. Outcomes12

[50] Alipanahi B, Delong A, Weirauch M T and Frey B J 2015 Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning Nat. Biotechnol.33

[51] Litjens G, Kooi T, Bejnordi B E, Setio A A A, Ciompi F, Ghafoorian M, van der Laak J A W M, van Ginneken B and Sánchez C I 2017 A survey on deep learning in medical image analysis Med. Image Anal.42

[52] Chauhan, A.S.; Lilhore, U.K.; Gupta, A.K.; Manoharan, P.; Garg, R.R.; Hajjej, F.; Keshta, I.; Raahemifar, K. Comparative Analysis of Supervised Machine and Deep Learning Algorithms for Kyphosis Disease Detection. Appl. Sci. 2023, 13, 5012. https://doi.org/10.3390/app13085012

[53] Chauhan, A. S.; Singh, J.; Kumar, S.; Saxena, N.; Gupta, M.; Verma, P. Design and Assessment of Improved Convolutional Neural Network Based Brain Tumor Segmentation and Classification System. J Integr Sci Technol 2024, 12, 793. https://doi.org/10.62110/sciencein.jist.2024.v12.793

[54] Naeem, A. B.; Senapati, B.; Chauhan, A. S.; Makhija, M.; Singh, A.; Gupta, M.; Tiwari, P. K.; AbdelRehim, W. M. F. . Hypothyroidism Disease Diagnosis by Using Machine Learning Algorithms. Int J Intell Syst Appl Eng 2023, 11, 368-373.

[55] Novita, M., Chauhan, A. S., Ujianti, R. M. D., Marlina, D., Kusumo, H., Anwar, M. T., Piasecki, M., & Brik, M. G. (2024, May). Exploring deep learning and machine learning for novel red phosphor materials. Journal of Luminescence, 269, 120476. https://doi.org/10.1016/j.jlumin.2024.120476